A Thesis Submitted for the Degree of Doctor of Informatics

Improvement of Sound Source Localization for a Binaural Robot of Spherical Head with Pinnae

by

Ui-Hyun KIM

Department of Intelligence Science and Technology

Graduate School of Informatics

Kyoto University

September 2013

Abstract

Based on binaural signals, i.e., the signals observed at the two ears, a listener can localize different sound sources in terms of three-dimensional position: the azimuth or horizontal angle, the elevation or vertical angle, and the distance for static sound sources or velocity for moving sound sources. For decades, researchers have tried to invent a robot that can do the same function under similar acoustic conditions. Despite all the efforts, the human auditory system is superior to any binaural robot audition system that has been devised so far. The topic of this thesis is to improve technical methods of sound source localization (SSL) for binaural robot audition. Since a binaural robot audition system consists of only two microphones embedded in the robot head, attaining high SSL performance is still difficult and not comparable to that with many microphones.

In this thesis, an improved SSL method is presented that is based on the generalized cross-correlation (GCC) method weighted by the phase transform (PHAT) for binaural robots equipped with a spherical head and two microphones inside human-like pinnae. In addition, the statistical model-based voice activity detection (VAD) algorithm employing the two-step noise reduction (TSNR) technique with recursive noise adaptation is also proposed to effectively reduce unexpected SSL errors during sound-absent periods. The conventional GCC-PHAT-based SSL method has four main problems when used on a binaural robot platform: 1) low-resolution time-delay-of-arrival (TDOA) estimation in the time domain, which makes SSL inaccurate in all directions and impossible in some cases, 2) diffraction of sound waves with multipath interference caused by the shape of the robot head, which degrades SSL accuracy mostly in the lateral directions, 3) front-back ambiguity, which limits the localization range to the front horizontal space, and 4) difficulties with multisource sound localization in real environments. The low-resolution problem is solved by using a maximum likelihood (ML)-based SSL method in the frequency domain. The

diffraction problem was overcome by incorporating a new time delay factor into the GCC-PHAT method under the assumption of a spherical robot head. The ambiguity problem was overcome by utilizing the amplification effect of the pinnae. Finally the difficulties with multisource sound localization in real environments were addressed by extending the proposed ML-based SSL method using the new time delay factor to enable simultaneous multiple direction estimations with signal-to-noise ratio (SNR)-weighting function and *K*-means clustering algorithm. For effective localization of an unknown time-varying number of multiple sound sources, the standard *K*-means clustering algorithm was improved by adding two additional steps that increase the number of clusters automatically and eliminate clusters containing incorrect direction estimations. Experiments conducted using a SIG-2 humanoid robot in single sound source situations showed that the proposed ML-based SSL method using the new time delay factor reduced the average localization error by 17.92° (2.23° vs. 20.15°) and the localization errors for the side directions by over 35° than the one using the conventional GCC-PHAT-based SSL method; the proposed disambiguation method using the pinna amplification effect had an average success rate with 22.5 points higher (92.28% vs. 69.78%) than the one using a conventional HRTF-based method. Experiments conducted in multisource sound situations also showed that the proposed multisource sound localization method could localize the unknown time-varying number of multiple sound sources in real time with localization errors below 5.96°.

This thesis consists of eight chapters. Chapter 1 shows the motivation and background of the thesis with five main problems in binaural sound localization and five approaches to the problem solving. Chapter 2 gives an overview of existing binaural robot audition systems and techniques in signal processing. The binaural localization cues and existing computational techniques in this chapter serves as the background in the subsequent chapters of the thesis. Chapter 3 summarizes the existing statistical model-based VAD algorithm and presents an improved VAD method using the TSNR technique with recursive noise adaptation. This proposed VAD method is used as a significant building block in the SSL system to reduce unexpected SSL errors by deactivating SSL process during speech-absent period. Chapter 4 presents an improved GCC-PHAT-based SSL method with the ML estimation in the frequency domain and the new time delay factor under the assumption of the spherical head for binaural robot audition. Chapter 5 presents a front-back disambiguation method based on the pinna

amplification effect for SSL over the entire azimuth. Chapter 6 presents a multisource sound localization method employing SNR-weighting function and *K*-means clustering. In this chapter, the standard *K*-means clustering algorithm was improved by adding two novel additional steps for localizing the unknown time-varying number of multiple sound sources in real environments. Chapter 7 discusses the contributions of this thesis and some ideas for future extensions. Finally, Chapter 8 concludes the thesis.

Acknowledgments

This work would not have been possible without the complete support by Okuno Laboratory of the Department of Intelligence Science and Technology in the Graduate School of Informatics at Kyoto University. First and foremost, I would like to express the deepest appreciation to my advisor, Prof. Hiroshi G. Okuno, for his excellent guidance, invaluable assistance, and sincere caring. His support helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor for my Ph.D. study.

Besides my advisor, I would like to thank my committee members, Prof. Tatsuya Kawahara and Prof. Akihiro Yamamoto. Without their insightful questions and comments this thesis would not have been completed.

My sincere gratitude also goes to my co-advisers, Prof. Tetsuya Ogata of Waseda University and Prof. Kazuhiro Nakadai of HONDA Research Institute Japan (HRI-JP). Prof. Tetsuya Ogata gave me warm encouragement and practical advices to carry out my research in the initial stages. Prof. Kazuhiro Nakadai supervised me with constructive comments and guidance during my internship at HRI-JP.

Discussions with Dr. Tall Takahashi, Dr. Shun Nishide, Dr. Katsutoshi Itoyama, Dr. Ryu Takeda, Dr. Takeshi Mizumoto, Mr. Takuma Otsuka, and everyone else in Okuno Laboratory had illuminated my research direction. I am also grateful to them for carefully reading, listening, and commenting on countless revisions in my writings and presentations so far. In addition, I appreciate the clerical support offered by Ms. Hiromi Okazaki in Okuno Laboratory.

I want to thank Dr. Randy Gomez of HRI-JP, who was always willing to help and give his best suggestions as a good friend.

Special thanks to Prof. Hyogon Kim of Korea University, Dr. Doik Kim of Korea Institute of Science and Technology (KIST), and Prof. Young-Woong Ko of Hallym University. Their tuition and research guidance helped me to go one step further when I

was in Korea.

I would also like to express my gratitude to the Japanese government (MEXT: Ministry of Education, Culture, Sports, Science, and Technology) for their financial support granted through predoctoral fellowship for three years.

Finally, I would particularly like to thank my parents, elder sister, and elder brother. They were always supporting and encouraging me with their best wishes.

Contents

List of Figures

List of Tables

CHAPTER 1
Introduction

1.1 Motivation and Background

Robots are artificial agents with capabilities of perception and action, usually electro-mechanical machines guided by computer programs and electronic circuitries in the physical world often referred by researchers as workspace. Their use had been generalized in the industrial need for factory jobs until only a few years ago and this has led to the modern robot development. Nowadays robots tend to be found in the most technologically advanced societies in such various domains as search and rescue, military zone, mine and bomb detection, scientific exploration, entertainment and hospital care [1]. These new domains of applications imply a closer interaction between humans and robots. The concept of closeness is to be taken in its full meaning, humans and robots share the workspace but also share goals in terms of task achievement [2]. With the advance in artificial intelligence (AI) and signal processing, robots start to have perceiving and understanding capabilities through cameras, microphones, and other active perception sensors equipped in their body. They are becoming more sophisticated and humanlike both in appearance and intelligence. A humanoid robot is one of such robots. However, despite robots are increasingly expected to have perceptual capabilities similar to those of human beings in the increasing demands for the close interaction with humans, they are still greatly lacking in capability, particularly in the functional ability of the auditory system through embedded microphones.

The auditory perception is sometimes regarded as a more important and preferential capability than other perceptions to robots, because other perception sensors embedded in their body can only feel a fraction of the world around them and are limited depending on situations, e.g., the camera has a limited field of view and is

hampered in darkness while the microphone to perceive sounds works in all directions regardless of the surrounding light [3]. As time goes by, robots will get into our lives more deeply than now and will be exposed to a vast variety of acoustical environments. The number and types of sound sources easily vary and there may be physical obstacles to disturb sound sources reaching the microphones embedded in their body. The environment may be highly dynamic with moving sources and objects. Robots will be expected to understand the acoustic scene even under strongly adverse conditions.

Among the various auditory functions required for robots, effective sound source localization (SSL) is a key to understanding the acoustic scene and achieving closer human-robot interaction (HRI). For example, this enables them to face the person speaking and signal to him/her that they are ready to listen, thereby appearing to express an interest in conversing with the person. A common approach to implementing SSL is to equip the robot with many microphones [4]–[10]. However, this causes several problems, including higher maintenance costs, the use of more computational power, and degradation of the human-like appearance. In addition, equipping robots with many microphones means that a general-purpose software interface can no longer be used due to the unique microphone array configuration for each robot. Humans are binaural, which means they have two sound inputs, i.e., two ears [11]. For robots to appear humanoid or to be perceived to be like human beings, it should also have two sound inputs, i.e., two microphones inside two artificial pinnae, one on each side of its head like human ears. A binaural audition method should not require excessive computing power. Binaural hardware and its software can also be easily ported to various kinds of robot platforms and embedded in information and communication technology (ICT) applications [12]. Moreover, research on binaural audition can contribute to understanding the human hearing mechanism [13]–[14]. The development of a binaural SSL method is thus particularly important for robots.

Extensive studies of SSL by a number of researchers have revealed perceptual clues. They include the interaural level difference (ILD), the interaural time difference (ITD), and the spectral distortion caused by various parts of the body (the pinnae, head, shoulders, torso, etc.) [15]–[16]. These clues are implicitly included in the head related transfer function (HRTF) [17]. The ITD, more commonly referred to as the time delay of arrival (TDOA), plays an important role in SSL; the sound signals arrive at each microphone at different times for directions. One of the most widely used SSL methods

based on the TDOA between binaural inputs is the generalized cross-correlation (GCC) method with phase transform (PHAT) weighting [18]–[19]. GCC-PHAT is one of the most successful formulations of GCC and performs well for a single sound source even in noisy and reverberant environments [20]–[21].

The use of a microphone array with many microphones has improved SSL performance on various robot platforms in actual environments. A reduction in the number of microphones generally degrades SSL performance. Since a binaural robot audition system uses only two microphones (one embedded on each side of the robot head), there are difficulties in obtaining performance as good as that with a microphone array. The main focus of this thesis is on improving methods for the localization of sound sources by only two sound inputs and minimum requirements. As the final outcome, the real-time binaural audition system proposed in this thesis requires only the diameter of robot head for SSL without any prior information such as impulse response data or learning parameters. These conditions may enable robot auditory systems to be simply and inexpensively implemented. In particular, to achieve an effective SSL system for binaural robot audition, this thesis addressed five problems with GCC-PHAT-based SSL on a binaural robot platform in real environments.

1.2 Problems in Binaural Sound Localization

The research objective in this thesis is to improve the localization performance by using only two microphones for the robots to be deployed in various acoustic environments. To attain this purpose, the robot audition system needed to localize a sound source with the 1° resolution over the entire azimuth and track multiple sound sources effectively in noisy and reverberant environments. This section introduces the five problems to be overcome, which are limiting or affecting the accuracy of SSL based on the GCC-PHAT method in binaural robot audition.

1.2.1 Problem 1: Voice Activity Detection with Insufficient SNR Estimation

Voice Activity Detection (VAD) is one of the most important auditory functions for robots because the target sound signals to be recognized by robots are usually human

speech. In addition, VAD can facilitate auditory processing in which the presence or absence of human speech is detected, and can also be used to deactivate SSL process during speech-absent period of a sound signal to reduce the computational cost and unexpected errors in SSL. Most VAD algorithms have the weakness that their performance is not good enough in a situation where the background noise level is high. The statistical model-based VAD algorithm is one of such VAD algorithms due to the insufficient signal-to-noise ratio (SNR) estimation [22]–[23]. To effectively reduce unexpected SSL errors in noisy environments, the statistical model-based VAD needs to provide the more accurate period of the target sound source to the SSL process even in the low SNR case.

1.2.2 Problem 2: Low-Resolution TDOA Estimation in Time Domain

Since the conventional GCC-PHAT method estimates TDOA as the sample delay in the time domain by the term of the inverse Fourier transform of the cross-power spectrum for its computational efficiency, the estimated TDOA for SSL is restricted to an integer value of the sample delay with the problem of low-resolution. This restriction makes SSL inaccurate with unequal resolution in all directions and impossible in some cases. Simple solutions to this problem involve widening the distance between two microphones and increasing the sampling frequency, but they still have inherent limitations in binaural robot audition with fixed head shape and processing power. To get accurate SSL performance with the 1° resolution, this problem of low-resolution TDOA estimation in the time domain needs to be solved by a different approach on the conventional GCC-PHAT method.

1.2.3 Problem 3: Diffraction of Sound Waves with Multipath Interference caused by Contours of Robot Head

Most SSL methods are based on the assumption that a microphone array is located in a free space environment; i.e., they do not take into account any diffraction of sound waves in non-free space environments like robot platforms. Sound waves easily bend and spread around the robot head, resulting in a difference in TDOA between the waves that travel around the front of the robot head and those that travel around the back of the

robot head. This diffraction of sound waves with multipath interference degrades localization performance of binaural robot audition, especially for sound sources in the lateral directions (around ±90°). This is because the diffraction of sound waves and multipath interference increase as the sound incidence goes to −90° or +90°. For accurate binaural SSL on the robot head, the diffraction of sound waves with multipath interference also needs to be considered on the conventional GCC-PHAT method.

1.2.4 Problem 4: Front-Back Ambiguity due to Same TDOA

Binaural audition methods localize a sound source as coming from the front despite the actual sound source being in the rear, because a sound source appears to be at equal (mirror) angles in the front and rear hemi-fields due to having the same TDOA for the front and back. This front-back ambiguity limits the localization range to the front horizontal space, from −90° to +90°. Current methods for solving the ambiguity problem involve using a microphone array with many microphones, using head movement [24]–[26], and using a specific HRTF database [27]–[28]. However, these methods have certain drawbacks: using a microphone array does not fit to the concept of binaural robot audition using two microphones; using head movement does not work well for short words or phrases, such as when someone calls the robot's name, because the robot needs enough time to complete its head movement [29]; using an HRTF database does not work if the system and environment change because its performance depends greatly on the system and environment [30]–[31]. For these reasons, the problem of front-back ambiguity needs to be overcome by a new approach to extend the localization range of binaural robot audition systems over the entire azimuth.

1.2.5 Problem 5: Difficulties with Multisource Sound Localization due to Correlated Sound Sources in Real Environments

Another significant problem to be overcome for SSL in binaural robot audition is the difficulty with multisource sound localization in real environments [32]–[33]. Localization performance generally drops as the number of microphones is reduced. Since a binaural robot audition system uses only two microphones embedded on each side of the robot head, the number of sound sources that a binaural robot audition

system can localize has been limited to a single source. The most commonly used multisource sound localization method is multiple signal classification (MUSIC) but it is unable to localize multiple sources with binaural sound inputs [34]. In addition, most binaural methods to localize multiple sound sources are based on their specific HRTFs with the limited localization resolution and their input sound signals has been assumed to be white noise or simple sinusoidal signals in the anechoic chamber with the ideal condition [35]–[36]. In real environments multiple sound sources to be localized for robots, e.g., human speech, are generally broadband signals which are easily correlated each other and corrupted by noise or reverberation as opposed to pure tones with the ideal condition. To achieve multisource sound localization and increase its localization accuracy in real environments, this practical problem with the correlated or noise-corrupted sound sources needs to be overcome.

1.3 Approaches to Problem Solving

This section introduces five solutions to the five problems described above that have effectively improved localization performance in the binaural robot audition with a pair of microphones. These five solutions took different approaches from the existing solutions, such as increasing the sampling frequency or the distance between two microphones, using HRTFs, using head movements, assuming white noise or pure tone signals as the input sound signals in the ideal condition, etc. This is because increasing the sampling frequency or the distance has their inherent limitations with fixed head shape and processing power, and the effectiveness of using HRTFs is highly dependent on its system configuration and environment. In addition, the HRTF database needs to be measured in 1° step for high-resolution localization. Moreover, the head movement causes self-motor noise for robots and the approach using white noise or pure tone signals as the input sound signals is impractical when the target sound signals are usually human speech to be recognized by robots.

These novel solutions developed by different approaches were implemented as a real-time sub-system in the “HARK” open-source robot audition software [37]–[39]. The resulting HARK-binaural system was evaluated experimentally using the SIG-2 humanoid robot.

1.3.1 Solution 1: Improved SNR Estimation using Two-Step Noise Reduction

The performance problem on the existing statistical model-based VAD algorithm is the insufficient *a priori* SNR estimation by the power subtraction method. The statistical model-based VAD algorithm simply uses the power subtraction method to estimate the *a priori* SNR, even though the estimated *a priori* SNR plays an important role in the VAD processing. In addition, it also assumes that the noise variance is already known through the noise statistic estimation procedure in advance despite the noise spectrum is changeable over time. These two weaknesses limit its VAD performance. The insufficient a *priori* SNR estimation on the existing statistical model-based VAD algorithm was improved by utilizing the two-step noise reduction (TSNR) technique [40]–[41] with recursive noise adaptation for effective SSL in the low SNR case.

1.3.2 Solution 2: Maximum-Likelihood Estimation in Frequency Domain

For the problem of low-resolution TDOA estimation in the time domain, the maximum likelihood (ML) estimation was applied to the GCC-PHAT method in the frequency domain. This enables direction estimations to be more accurate with the 1° resolution than those of the conventional way calculated by the term of the inverse Fourier transform in the cross-power spectrum phase (CSP) analysis [42]. In this ML-based GCC-PHAT method, it was assumed that frequency bins of the cross-power spectrum are obtained from the exponential distribution and the true sound incidence direction is estimated by finding a parameter value of the distribution in the ML estimation that maximizes the sum of the cross-power spectrum with PHAT weighting in the frequency domain without the term of the inverse Fourier transform.

1.3.3 Solution 3: New Time Delay Factor

Under the assumption that the robot head is spherical for general shape of robot heads, a new time delay factor was derived to compensate for the diffraction of sound waves along with the robot head and multipath interference between the waves that travel around the front of the robot head and those that travel around the back of the robot

head. This new time delay factor was incorporated into the GCC-PHAT method instead of the conventional time delay factor derived in a free space condition. The ML-based GCC-PHAT method incorporated with the new time delay factor aids in estimating sound directions more accurately on the spherical robot head.

1.3.4 Solution 4: Front-Back Disambiguation by using Amplification Effect of Pinnae

To extend the localization range of binaural robot audition systems over the entire azimuth, a new decision rule for front-back disambiguation was derived based on the pinna amplification effect that creates a level difference between sound signals coming from the front and back. If the observed signal is determined to be behind the robot head equipped with silicon human-like pinnae by using this decision rule, the sound direction estimated by the SSL method in the front horizontal space is switched to the mirrored angle location in the back for SSL over the entire azimuth.

1.3.5 Solution 5: SNR-Weighting Function and Improved *K*-means Clustering

The solution to realizing multisource sound localization is threefold: the GCC-PHAT method was extended to enable simultaneous multidirection estimations for each time frame. Then a SNR-weighting function was incorporated into the GCC-PHAT method to cope with noise-corrupted sound sources. Finally the *K*-means clustering [43] was performed in order to eliminate incorrect direction estimations caused by the problem of correlation between sound sources in real environments.

1.4 Thesis Organization

This thesis consists of eight chapters. Its organization and conceptual diagrams from SSL to HRI is outlined in Fig. 1.1.

Chapter 2 surveys the literature related to robot audition and signal processing. The binaural localization cues and existing computational techniques in this chapter serves as the motivation and background for the works in the subsequent chapters.

Chapter 3 summarizes the existing statistical model-based VAD algorithm that still has the problem of insufficient *a priori* SNR estimation and presents an improved VAD method using the TSNR technique and recursive noise adaptation as the solution. This proposed VAD method is used as a significant building block in the binaural sound localization system to reduce unexpected SSL errors by deactivating SSL process during speech-absent period.

Chapter 4 presents an improved SSL method based on the GCC-PHAT method for binaural robot audition. The two problems with the conventional GCC-PHAT-based SSL method are the low-resolution TDOA estimation and the diffraction of sound waves with multipath interference caused by the shape of the robot head. These two problems affect the localization accuracy in binaural robot audition. The problem of low resolution was solved by applying the ML estimation in frequency domain and the diffraction problem was overcome by incorporating the new time delay factor into the GCC-PHAT method under the assumption of the spherical robot head. Experimental results are presented and the improvements of localization accuracy are evaluated with discussions in a single sound source situation.

Chapter 5 describes a binaural sound localization method over the entire azimuth for use with robots equipped with two microphones inside artificial pinnae. The problem of front-back ambiguity, which limits the localization range to the front horizontal space, was overcome by utilizing the amplification effect of the pinnae as a novel solution for localization range over the entire azimuth. Experimental results in SSL over the entire azimuth and in front-back disambiguation are presents and evaluated.

Chapter 6 presents a multisource sound localization method employing SNR-weighting function and *K*-means clustering. The ML-based GCC-PHAT method was extended to enable simultaneous multiple direction estimations with the SNR-weighting function. The standard *K*-means clustering algorithm was improved by adding two novel additional steps for effective multisource sound localization in real environments. The experimental results changing SNR in the twelve different conditions of two- and three-speakers are presented and evaluated with discussions.

Chapter 7 presents observations, contributions, and some ideas for future extensions.

Finally, Chapter 8 draws the conclusion.

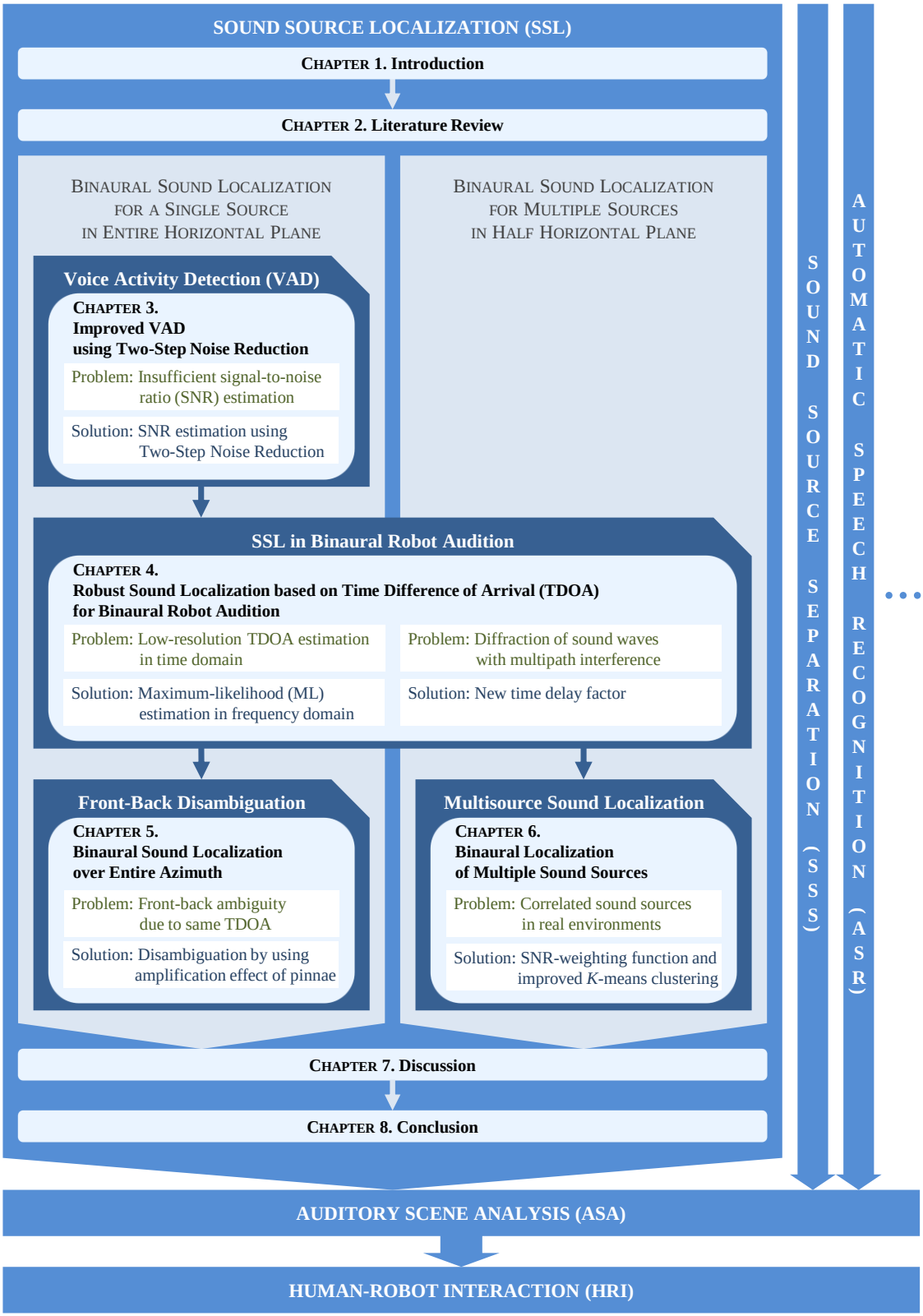

Figure 1.1: Thesis organization.

CHAPTER 2
Literature Review

2.1 Related Works to Robot Audition

This section gives an overview of existing binaural robot audition systems for SSL and the HARK open-sourced robot audition software.

2.1.1 Researches on Humanoid Robots

Although many researchers majoring in digital sound processing have developed robot audition systems, a few binaural robot audition systems had been developed so far. Two binaural audition systems achieved on the SIG and SIG-2 humanoid robots are representative.

Nakadai et al. developed a talker tracking system for the front horizontal space on the SIG humanoid robot [44]–[48]. The SIG humanoid robot is an upper torso humanoid and it has a plastic cover designed to acoustically separate its interior from the external world. It is fitted with a pair of cameras for stereo vision and two pairs of microphones for auditory processing: one in the left and right ears for collecting sounds from the external world, and the other one inside the cover mainly for canceling self-motor noise in motion. The robot audition system on the SIG humanoid robot localized multiple speakers by using HRTF database approximated by the scattering theory [49] as shown in Fig. 2.1. Since HRTF is usually measured with 10° or 15° steps in an anechoic chamber to reduce the cost of measuring HRTFs and therefore the system could only direction estimations stored in the HRTF database. Thus they approximated HRTF by using geometrical relation and scattering theory for SSL with high localization resolution. In addition, their active direction-pass filter (ADPF) could separate sound sources by

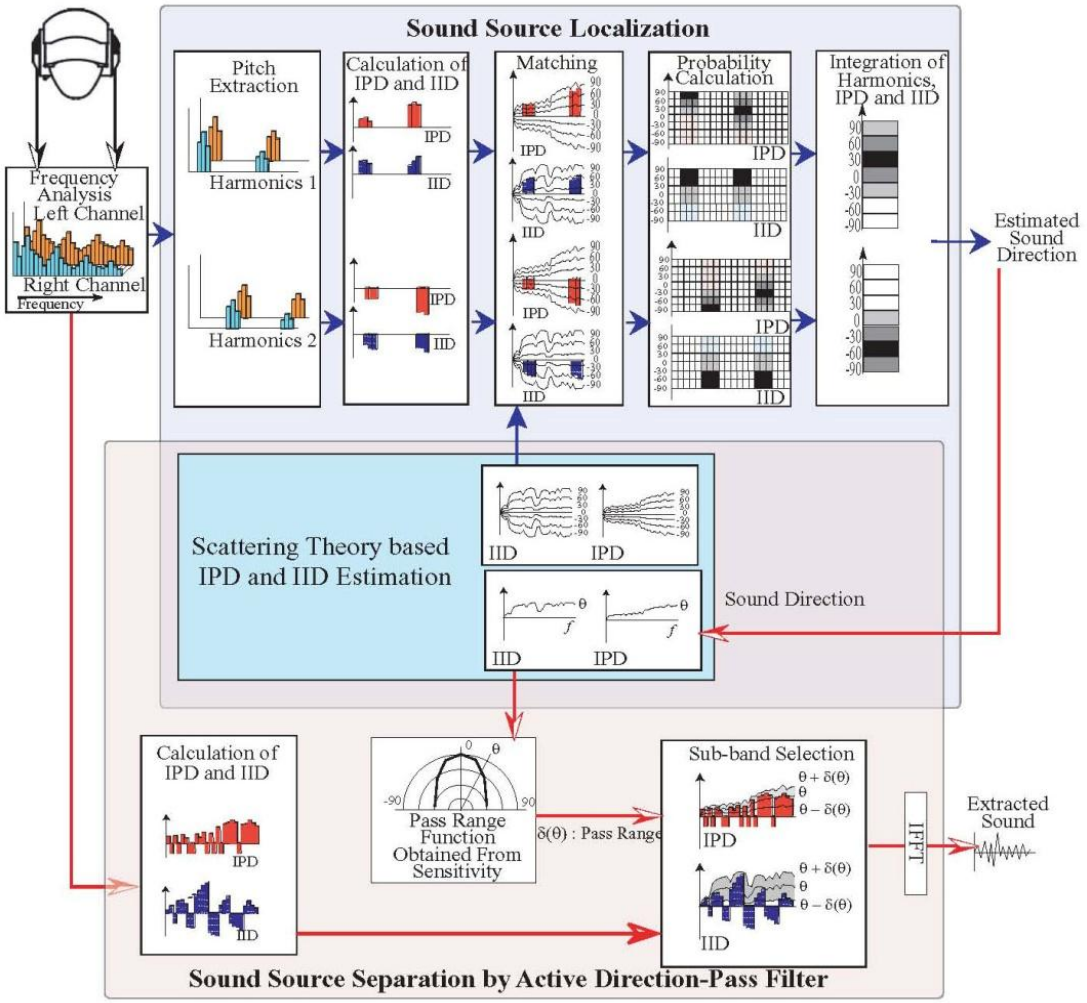

Figure 2.1: Sound localization and separation system using scattering theory for SIG humanoid robot [47].

emphasizing the target signals on sub-bands where the interaural phase difference (IPD) and interaural intensity difference (IID), i.e., ILD, match those of the specific sound directions. However, the performance of this system is highly dependent on its system configuration and environment with the problem of low localization resolution due to the fundamental method using HRTF database and a lot of setting parameters required by the scattering theory.

Kim et al. developed a binaural active audition system on the SIG-2 humanoid robot with movements [50]–[57]. The SIG-2 humanoid robot was designed so as to solve

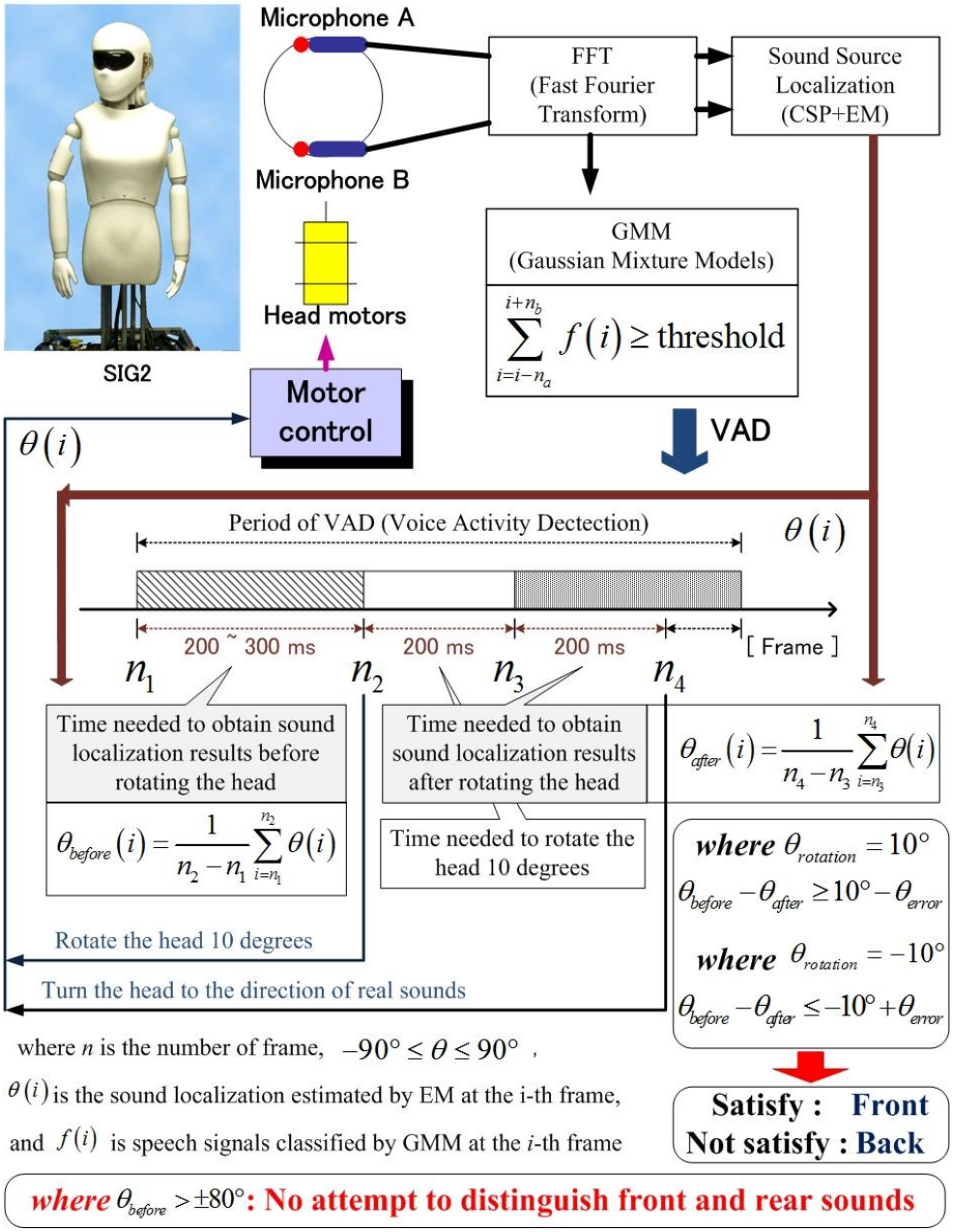

Figure 2.2: Sound localization system based on CSP analysis and EM algorithm using head movements for SIG-2 humanoid robot [57].

some problems in the SIG humanoid robot, such as the loud self-motor noise, a strong sound reflection by the body, sound resonance and leakage inside the cover, and lack of pinnae. The SIG-2 humanoid robot has two cameras for stereo vision and two microphones inside the silicon human-like pinnae for auditory processing. For the SIG-2

humanoid robot, the robot audition system was realized by using the CSP analysis based on the ITD information and the expectation-maximization (EM) algorithm to filter SSL errors and compensate for the estimated directions in real time. In addition, the head movement was utilized for front-back disambiguation: if a sound is coming from the front, the sound direction will be decreased while turning the robot head to the sound direction whereas the sound direction will be increased if s sound is coming from the back. Figure 2.2 shows the process of this auditory system on the SIG-2 humanoid robot to localize sound sources over the entire azimuth. First, the SIG-2 humanoid robot detects speech signals by using the VAD based on the Gaussian mixture model (GMM) and tracks a speech source by using the CSP analysis and the EM algorithm. Then the system starts to turn its head 10° towards the direction of the detected sound to distinguish whether the detected sound is in the front or the back. Finally, the SIG-2 humanoid robot localizes sounds over the entire azimuth. However, the CSP analysis has the problem of low-resolution TDOA estimation in the time domain that will be addressed in more detail in this thesis. In addition, the head movement causes self-motor noise and requires a sound source persisting until the robot completes the head movement.

2.1.2 Robot Audition Software: HARK

HARK (HRI-JP audition for robots with Kyoto University), an ancient English word for “listen”, is open-sourced robot audition software consisting of SSL modules, sound source separation (SSS) modules, and automatic speech recognition (ASR) modules of separated speech signals that works on any robot with any microphone configuration [58]. Under the purpose to give robots an easy way to equipped with auditory functions to cope with various auditory environments, HARK provides gives a comprehensive set of functions enabling computational auditory scene analysis (ASA) [59] from any robot, any microphone configuration, and any hardware. The resulting implementation of HARK with MUSIC-based sound source localization, geometric high-order dicorrelation-based source separation (GHDSS)-based sound source separation, and missing feature theory (MFT)-based automatic speech recognition (ASR) attains estimating and recognizing multiple sound sources in real time [60]–[61]. Within the recent revival of binaural audition system, HARK has been complemented with an

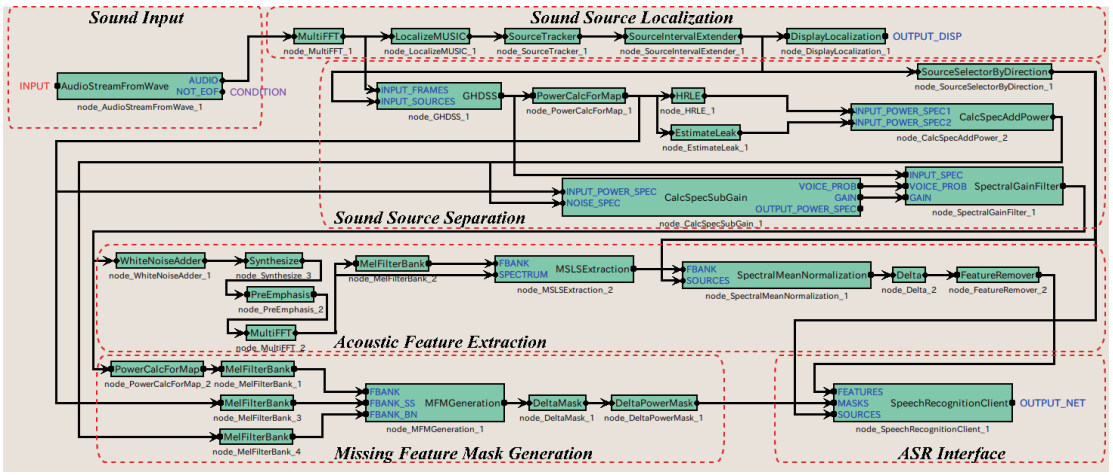

Figure 2.3: Module construction on FlowDesigner for typical robot audition with HARK [58].

unofficial package for binaural sound localization processing but this package requires lots of prior information such as impulse response data and parameter setting that is sensitive to auditory environments.

HARK also provides a set of graphical modules to be programed in C++ by users through an open-sourced middleware, FlowDesigner. The main feature of FlowDesigner is that user can construct a system using a graphical user interface [62]. The programming in FlowDesigner is achieved by laying out and connecting nodes, and setting property values of the nodes as shown in Fig. 2.3. The proposed SSL methods with the solutions throughout this thesis were implemented as real-time sub-systems in HARK and evaluated experimentally using the SIG-2 humanoid robot.

2.2 Related Works to Signal Processing

This section introduces the primary clues for binaural SSL. The primary clues include the ILD, the ITD, the IPD, and the spectral modification mostly caused by the pinnae. These clues are implicitly included in HRTF as shown in Fig. 2.4. In addition, this section also introduces the beamforming technique, the MUSIC algorithm, the GCC method, and the time-frequency representation that are widely used in the audio signal processing field.

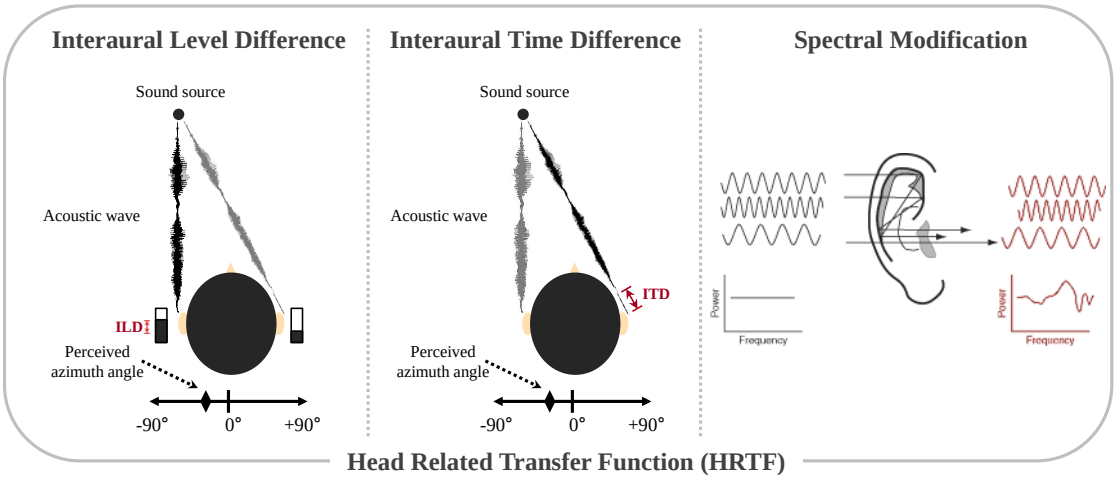

Figure 2.4: Primary clues for binaural SSL.

2.2.1 Binaural Sound Localization Cues

The reason that we can localize the source of a sound accurately is that we have two ears. At each ear, a slightly different signal will be perceived and by analyzing the differences between two ears the brain can determine where the sound originated. The most important localization cues are ILD and ITD. The ILD arises from the fact that, due to the shadowing of the sound wave by the head (head shadow) [63]–[64], a sound coming from a source located to one side of the head will have a higher intensity, or be louder, at the ear nearest the sound source. The ITD occurs whenever the distance from the source of sound to the two ears is different, resulting in differences in the arrival times of the sound at the two ears. Sound from the right side reaches the right ear earlier than the left ear. When the sound source is directly in front of the head, there is no ITD, i.e., the ITD is zero. ITD changes systematically with azimuth. The IPD refers to the difference in the phase of a wave that reaches each ear, and is dependent on the frequency of the sound wave and ITD in the frequency domain. Imagine a 1000 Hz tone that reaches the left ear 0.5 ms before the right. As the wavelength reaches the right ear, it will be 180° out of phase with the wave at the left ear. Once the brain has analyzed ILD, ITD, and IPD, the location of the sound source can be determined with relative accuracy. ITDs and IPDs are mainly evaluated for frequencies below 800 Hz and ILDs

are mainly evaluated for frequencies above 1600 Hz. Between 800 Hz and 1600 Hz there is a transition zone, where these mechanisms, ILD, ITD, and IPD, play a role. It is known as the human ear has the ability to detect differences as small as 1° for sound sources in front and 15° for sound sources to the sides [65]–[66].

The external ear provides a frequency and direction dependent filter of incoming sound. These spectral cues arise from path differences in sounds reflected mostly from the pinna. Due to the principle of superposition and the asymmetry of the pinnae, the sounds of different frequencies will produce a resonance in the ear canal which differs with sound source location. The path differences involved are small and direction dependent filtering will be greatest for high frequency sounds at which a small path difference causes a large phase difference. Significant spectral features will therefore occur at higher frequencies for animals with smaller ears. At these frequencies, one finds a complicated pattern of peaks and notches which varies with sound source direction. It has been suggested that the brain also uses these features as cues to SSL. Spectral cues help to resolve front-back ambiguity. For a spherical head without external ear structures, sound sources in front and behind will ideally have the same ITD and the same ILD. The spectral filtering of the pinnae distorts ILDs across frequencies, resolving front-back ambiguity especially for broad-band sounds. This is the basis for localizing sounds in the vertical plane. For example, assuming the two ears are symmetrical, sounds presented from different elevations along the anterior mid-sagittal plane will have identical ILDs and ITDs but will differ in their spectral contents [67].

2.2.2 Head-Related Transfer Function

A HRTF is a response that characterizes how an ear receives a sound from a point in space; a pair of HRTFs for two ears can be used to synthesize a binaural sound that seems to come from a particular point in space. It is a transfer function, describing how a sound from a specific point will arrive at the ear. Humans have just two ears, but can locate sounds in three dimensions–in distance, in direction of elevation and azimuth. This is possible because the brain, the inner ear and the external ears (pinnae) work together to make inferences about location. Humans estimate the location of a source by taking cues derived from one ear (monaural cues), and by comparing cues received at

both ears (binaural cues). The binaural cues are mainly ILD and ITD. The monaural cues come from the interaction between the sound source and the human anatomy, in which the original source sound is modified before it enters the ear canal for processing by the auditory system. These spectral modifications encode the source location, and may be captured via an impulse response which relates the source location and the ear location. This impulse response is termed the head-related impulse response (HRIR) and the HRTF is the Fourier transform of HRIR. The HRTF describes how a given sound wave input is filtered by the diffraction and reflection properties of the head, pinna, and torso, before the sound reaches the transduction machinery of the eardrum and inner ear.

Linear systems analysis defines the transfer function as the complex ratio between the output signal spectrum and the input signal spectrum as a function of frequency. Therefore, the HRTF $H[f, \theta]$ of any dummy head system at frequency f in sound direction θ can be technically obtained from the synchronized input sound spectrum and the output sound spectrum observed at each microphone installed in the dummy head as follows:

$$H[f,\theta] = \frac{\text{Output}[f,\theta]}{\text{Input}[f,\theta]}. \tag{2.1}$$

HRTFs are typically measured in an anechoic chamber to minimize the influence of early reflections and reverberation on the measured response. HRTFs are usually measured at increments of θ such as 10° or 15° in the horizontal plane as a lot of essential prior data for SSL [68].

2.2.3 Beamforming

Beamforming (BF) is a spatial filtering technique used in microphone arrays for directional signal transmission [69]. It is based on ITD and therefore strongly related to the classical ITD approach. The main deference to ITD is that there is no exact calculation but position estimation by directing the beamformer through space and looking for the maximal output. Directing the beamformer and looking for the highest output simply can be done with a conventional beamformer that is also known as the delay-and-sum beamformer. In the delay-and-sum beamformer all the weights of the microphone elements can have equal magnitudes and it is steered to a specified direction only by selecting appropriate phases for each microphone [70]. If the noise is uncorrelated and

there are no directional interferences, the output power of the delay-and-sum beamformer with M microphones receiving an input signal of power X can be defined as:

$$\hat{P}_{BF}\left(e^{j(m-1)\omega}\right)=\sum_{m=1}^{M}\frac{X_m}{\sigma^2_{noise}}e^{j(m-1)\omega}, \tag{2.2}$$

where m is the index of microphones, σ^2_{noise} is noise variance or noise power to whiten the target signal power, and e is an steering direction vector in phase to look for the highest output power.

The main problem using the delay-and-sum beamformer is that it has relatively wide energy peaks (beamwidth). This makes localization resolution poorer and therefore more difficult to localize target sources from other sources or noise, particularly with the binaural audition system using two microphones. These wide energy peaks can be narrowed by using multiple microphones arranged as an array system with the incensements of the robustness to noise and stability.

2.2.4 Multiple Signal Classification

MUSIC is an algorithm used for multisource sound localization with the microphone array. MUSIC estimates the frequency content of a signal or autocorrelation matrix using an eigenspace method [71]. This method assumes that a sound signal consists of F complex exponentials in the presence of Gaussian white noise. Given an M microphone sound inputs and their $M \times M$ autocorrelation matrix, R_{xx}, if the eigenvalues are sorted in decreasing order, the eigenvectors corresponding to the l largest eigenvalues, i.e., directions of largest variability, span the signal subspace. The remaining $M-l$ eigenvectors span the orthogonal space, where there is only noise.

The frequency estimation function for MUSIC is

$$\hat{P}_{MUSIC}\left(e^{j\omega}\right)=\frac{1}{\sum_{m=l+1}^{M}\left|e^H v_m\right|^2}, \tag{2.3}$$

where v_m are the noise eigenvectors and

$$e=\begin{bmatrix}1 & e^{j\omega} & e^{j2\omega} & \cdots & e^{j(M-1)\omega}\end{bmatrix}^T \tag{2.4}$$

is an arbitrary direction vector for SSL, which searches the whole direction area of interest. If *e* matches the true direction vectors during the searching, the denominator in (2.3) goes to zero, resulting in spatial spectrum peaks. MUSIC outperforms simple methods such as picking peaks of spatial spectra in the presence of noise, when the number of components is known in advance, because it exploits knowledge of this number to ignore the noise in its final calculation. In other words, its chief disadvantages are that it requires the exact number of sound sources to be known in advance for SSL and it has poor performance with the small number of microphones in real environments including noise and reverberation. Therefore it cannot be used in more general cases for sound processing, particular in binaural sound localization.

2.2.5 Generalized Cross-Correlation

In signal processing, cross-correlation (CC) is a measure of similarity of two waveforms as a function of a time-lag applied to one of them. This is also known as a sliding dot product or sliding inner-product. It is commonly used for TDOA estimation between binaural sound inputs for SSL. The GCC takes a weighting function into CC [72]–[74]. Various algorithms in this family can be defined by using different weighting functions, some of which depend on the specific observations and others on their statistical properties or estimates thereof. Optimal localization results are possible in this framework when the signals and noise or reverberation fit certain models. Other weighting functions work well empirically in many situations, one of which is the PHAT, in which the weighting factor cancels the magnitudes of the left and right signals to preserve only the phase information on CC [75].

The GCC-PHAT method does not require prior information, such as impulse response data or learning parameters, and can accurately find the direction of a sound in noisy and reverberant environments even with two microphones compared to the BF-based or MUSIC-based method. Therefore, the GCC-PHAT method was used and improved for binaural SSL in this thesis. The GCC-PHAT method can usually estimate one direction of sound source at a frame. For multisource sound localization, the GCC-PHAT method was extended to enable multidirection estimations with a SNR-weighting function and an improved *K*-means clustering method. Chapter 4 describes the conventional GCC-PHAT method in detail.

2.2.6 Time-Frequency Representation

In computational models for the analysis of audio signals in time and in frequency, the short-time Fourier transform (STFT) is the most commonly used signal representation. In practical implementations the windowed discrete Fourier transform (DFT), the discrete equivalent of the STFT, is applied to cope with the frequency and phase content of local sections of sound signals to change over time.

Simply, in the continuous-time case, the function to be transformed is multiplied by a window function which is nonzero for only a short period of time. The Fourier transform of the resulting signal is taken as the window is slid along the time axis, resulting in a time-frequency representation of the signal. Mathematically, the STFT is written as:

$$X[f,n] = \sum_{l} \omega_a[l]x[l+nO]\exp\left(-j2\pi\frac{l}{L}f\right), \tag{2.5}$$

where $\omega_a[l]$ is the analysis window function, commonly a Hann and Hamming window, of length L with time index, l, and $x[l]$ is the general signal to be transformed. $X[f,n]$ is essentially the Fourier Transform of $\omega[l]x[l+nH]$, a complex function representing the phase and magnitude of the signal over frequency, f, and time, n. The time index is related to the hop-size, O, that describes how much the window moves (in samples) between two consecutive time indexes, n and l.

The STFT is invertible, that is, the original signal can be recovered from the transform by the Inverse STFT (ISTFT) [76]. The most widely accepted way of inverting the STFT is by using the overlap-add (OLA) method, which also allows for modifications to the STFT complex spectrum [77]. Using a synthesis window function, $\omega_s[l]$, of equal length L, the signal can be reconstructed by the inverse DFT and the overlap-add technique. The Inverse STFT is defined as

$$x[l] = \sum_{n} \omega_s[l-nO]\sum_{f=1}^{L} X[f,n]\exp\left(j2\pi\frac{l}{L}f\right). \tag{2.6}$$

One of the downfalls of the STFT is that it has a fixed resolution. The length of the windowing function relates to how the signal is represented. It determines whether there is good frequency resolution (frequency components close together can be separated) or good time resolution (the time at which frequencies change). A wide

window gives better frequency resolution but poor time resolution. A narrower window gives good time resolution but poor frequency resolution. These are called narrowband and wideband transforms, respectively.

CHAPTER 3
Improved Voice Activity Detection

3.1 Introduction

VAD is an essential technique in the robot audition system. VAD can facilitate robot speech processing in which the presence or absence of human speech is detected, and can also be used to deactivate SSL process during speech-absent period of an audio signal to reduce the computational cost and unexpected SSL errors. The purpose of the VAD is to provide delimiters for the beginning and end of a continuous speech-present period as exactly as possible from background noise such as music or other non-speech signals. For this purpose, it first extracts some features or quantities from the audio signal and compares these observed values with those of estimated noise according to some decision rules [78].

The most conventional VAD algorithms are based on zero-crossing rate, periodicity estimation, and signal energy level detection. The most well-known algorithm of this kind is the G.729B VAD [79]–[80]. These conventional VAD algorithms have the weakness that their performance is not good enough in a situation where the background noise level is high. In other words, they cannot work well in the low SNR case. To cope with this problem, many improved VAD algorithms have been designed and proposed but they also have their drawback of using heuristics which makes it difficult to optimize the relevant parameters. Sohn et al. have proposed a VAD based on a statistical model. This statistical model-based VAD algorithm requires fewer parameters for optimization than the G.729B VAD and uses the log likelihood ratio (LLR) of speech and background noise variances of statistics for the low SNR case [81].

However, this statistical model-based VAD algorithm simply uses the power subtraction method to estimate the *a priori* SNR, even though the estimated *a priori* SNR plays an important role in the VAD processing. In addition, it also assumes that the noise variance is already known through the noise statistic estimation procedure in advance despite the noise spectrum is changeable over time. These two weaknesses thus limit its VAD performance:

- **Insufficient a *priori* SNR estimation**: the performance problem on the existing statistical model-based VAD algorithm is the insufficient a *priori* SNR estimation by the power subtraction method and the absence of noise adaptation.

To provide more accurate VAD results for effective SSL in the low SNR case, this insufficient a *priori* SNR estimation needs to be improved:

- **TSNR technique with recursive noise adaptation**: the insufficient a *priori* SNR estimation on the existing statistical model-based VAD algorithm was improved by utilizing the noise reduction technique and recursive noise adaptation instead of the use of the power subtraction method.

The chapter is organized as follows. Section 3.2 summarizes the existing statistical model-based VAD algorithm. Section 3.3 presents the improved statistical model-based VAD algorithm employing the TSNR technique with recursive noise adaptation. Section 3.4 evaluates experimental results with discussions. Finally, Section 3.5 concludes this chapter.

3.2 Statistical Model-based Voice Activity Detection Algorithm

Assuming that clean speech is degraded by uncorrelated additive noise, the observed signal can be represented with two hypotheses, speech absence H_0 and speech presence H_1, as follows:

$$H_0 : speech\ absent\ : X[f,n] = N[f,n] \tag{3.1}$$

$$H_1 : speech\ present : X[f,n] = S[f,n] + N[f,n]. \tag{3.2}$$

where $X[f,n]$, $S[f,n]$, and $N[f,n]$ are f-th elements of the STFT of the noisy speech, clean speech, and uncorrelated additive noise, respectively, on the n-th time-frame index in the

time-frequency representation. The $f \in \{1, \ldots, F\}$ denotes a frequency bin, F is the frame size of the STFT , fs is the sampling frequency.

Adapting the Gaussian statistical model that means the STFT coefficients of clean speech and uncorrelated additive noise are asymptotically independent Gaussian random variables, the probability density functions (PDF) conditioned on two hypotheses H_0 and H_1 are given by

$$p(X[f,n]|H_0) = \prod_{f=0}^{T-1} \frac{1}{\pi\lambda_N[f,n]} \exp\left\{-\frac{|X[f,n]|^2}{\lambda_N[f,n]}\right\}, \tag{3.3}$$

$$p(X[f,n]|H_1) = \prod_{f=0}^{T-1} \frac{1}{\pi\left(\lambda_S[f,n]+\lambda_N[f,n]\right)} \cdot \exp\left\{-\frac{|X[f,n]|^2}{\lambda_S[f,n]+\lambda_N[f,n]}\right\}, \tag{3.4}$$

where $\lambda_S[f,n]$ and $\lambda_N[f,n]$ is the variances of $S[f,n]$ and $N[f,n]$, respectively. Based on the assumed statistical models, the likelihood ratio (LR) is

$$\Lambda[f,n] = \frac{p(X[f,n]|H_1)}{p(X[f,n]|H_0)} = \frac{1}{1+\xi[f,n]} \exp\left\{\frac{\gamma[f,n]\xi[f,n]}{1+\xi[f,n]}\right\}, \tag{3.5}$$

where $\xi[f,n]=\lambda_S[f,n]/\lambda_N[f,n]$ is the *a priori* SNR and $\gamma[f,n]=|X[f,n]|^2/\lambda_N[f,n]$ is the *a posteriori* SNR.

The *a priori* SNR $\xi[f,n]$ and the noise variance $\lambda_N[f,n]$ are unknown in advance as the noisy speech $X[f,n]$ alone is available. Therefore, $\xi[f,n]$ and $\lambda_N[f,n]$ need to be estimated by some procedure. This statistical model-based VAD algorithm assumes that $\lambda_N[f,n]$ is already known through the noise statistic estimation procedure and $\xi[f,n]$ can be derived by the power subtraction method as follows:

$$\hat{\xi}[f,n] = \frac{|X[f,n]|^2 - \lambda_N[f,n]}{\lambda_N[f,n]} = \gamma[f,n] - 1. \tag{3.6}$$

The VAD decision rule is derived from substituting Equation (3.6) into Equation (3.5) and the mean of the LLR for individual frequency bins:

$$\begin{aligned} &\text{if} \quad \hat{\Lambda}[n] > \eta_{VAD} \quad \text{then} \quad n = \text{speech present} \\ &\text{else} \qquad\qquad\qquad\quad n = \text{speech absent} \end{aligned} \tag{3.7}$$

where η_{VAD} is a threshold and

$$\begin{aligned} \hat{\Lambda}[n] &= \frac{1}{T}\sum_{f=0}^{f_s(T-1)/T} \log \Lambda[f,n] = \frac{1}{T}\sum_{f=0}^{f_s(T-1)/T} \log\left\{\frac{1}{\gamma[f,n]}\exp\big(\gamma[f,n]-1\big)\right\} \\ &= \frac{1}{T}\sum_{f=0}^{f_s(T-1)/T} \{\gamma[f,n]-\log\gamma[f,n]-1\}. \end{aligned} \tag{3.8}$$

3.3 Proposed Voice Activity Detection

This section gives the improved statistical model based VAD algorithm employing the TSNR technique with recursive noise adaptation. The TSNR technique is utilized to improve the insufficient *a priori* SNR estimation instead of the power subtraction method. Figure 3.1 shows the block diagram of the proposed VAD system. In the proposed VAD system, the *a priori* SNR is optimized after the TSNR technique with recursive noise adaptation, and then it is used in the existing statistical model-based VAD process which is described above.

3.3.1 *A priori* SNR Estimation using Two-Step Noise Reduction Technique

In the first step in the TSNR technique, the *a priori* SNR is computed with the decision-directed (DD) estimation approach to reduce the bias of an estimator [82]–[83] as follows:

$$\hat{\xi}_{DD}[f,n] = \alpha\frac{\left|\hat{S}[f,n]\right|^2}{\lambda_N[f,n]} + (1-\alpha)P\{\gamma[f,n]-1\}, \tag{3.9}$$

where $P[\cdot]$ is the half-wave rectification which is defined by $P[x]=x$ if $x\geq 0$ and $P[x]=0$ otherwise, and α is the forgetting factor whose value is typically chosen as 0.98 ($0 < \alpha <$

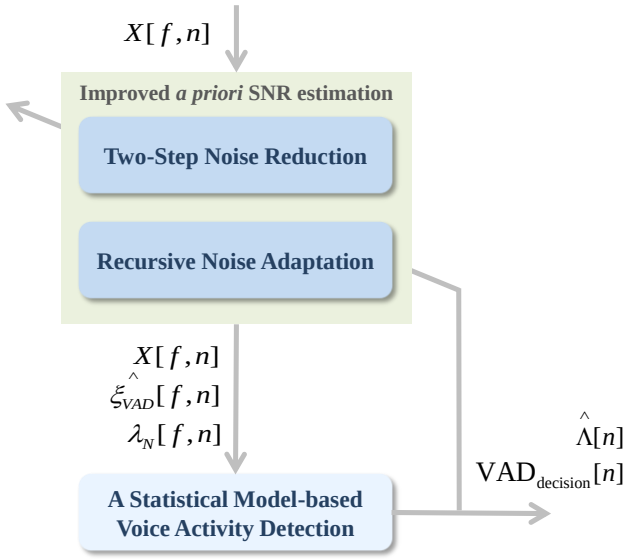

Figure 3.1: Block diagram of an improved statistical model-based VAD algorithm employing the TSNR technique and recursive noise adaptation.

1). Then, the spectral gain $G_{DD}[f,n]$ is obtained by applying (3.9) to the Wiener amplitude estimator as follows:

$$G_{DD}[f,n] = \frac{\hat{\xi}_{DD}[f,n]}{1+\hat{\xi}_{DD}[f,n]}. \tag{3.10}$$

In the second step, $G_{DD}[f,n]$ is used for estimation of the TSNR *a priori* SNR as follows:

$$\hat{\xi}_{TSNR}[f,n] = \frac{|G_{DD}[f,n]X[f,n]|^2}{\lambda_N[f,n]}. \tag{3.11}$$

Finally, the spectral gain $G_{TSNR}[f,n]$ to enhance speech is obtained by applying (3.11) to the Wiener amplitude estimator again:

$$G_{TSNR}[f,n] = \frac{\hat{\xi}_{TSNR}[f,n]}{1+\hat{\xi}_{TSNR}[f,n]}, \tag{3.12}$$

and the enhanced speech can be obtained by applying $G_{TSNR}[f,n]$ to the noisy signal as the following equation:

$$\hat{S}_{TSNR}[f,n] = G_{TSNR}[f,n]X[f,n]. \tag{3.13}$$

3.3.2 Noise Adaptation

To compensate for fluctuations of noise power level, the noise variance $\lambda_N[f,n]$ is updated in a recursive way [83]–[84]as follows:

$$\lambda_N[f,n] = \beta_{VAD}\lambda_N[f,n-1] + (1-\beta)\left\{|X[f,n]|^2 - |\hat{S}_{TSNR}[f,n]|^2\right\}, \tag{3.14}$$

where β_{VAD} is the forgetting factor ($0 < \beta_{VAD} < 1$). This noise adaptation function is performed on the speech-absent frames determined by the VAD decision rule.

3.3.3 Improved Statistical Model-Based VAD Algorithm

The speech spectrum enhanced by the TSNR technique in (3.13) is utilized to improve the *a priori* SNR estimation for the statistical model-based VAD algorithm instead of the existing power subtraction method in (3.6) as follows:

$$\hat{\xi}_{VAD}[f,n] = \frac{\hat{S}_{TSNR}[f,n]}{\lambda_N[f,n]}. \tag{3.15}$$

An improved VAD decision rule to be substituted for (3.8) can be derived by substituting (3.15) into (3.5) and the mean of the LLR:

$$\hat{\Lambda}[n] = \sum_{f=0}^{f_s(T-1)/T} \log\left\{\frac{1}{1+\hat{\xi}_{VAD}[f,n]}\right\} + \frac{\gamma[f,n]\hat{\xi}_{VAD}[f,n]}{1+\hat{\xi}_{VAD}[f,n]}, \tag{3.16}$$

where the noise variance is recursively updated by (3.14).

3.4 Evaluation

The improved statistical model-based VAD system employing the TSNR technique with recursive noise adaptation was constructed and evaluated in the binaural audition

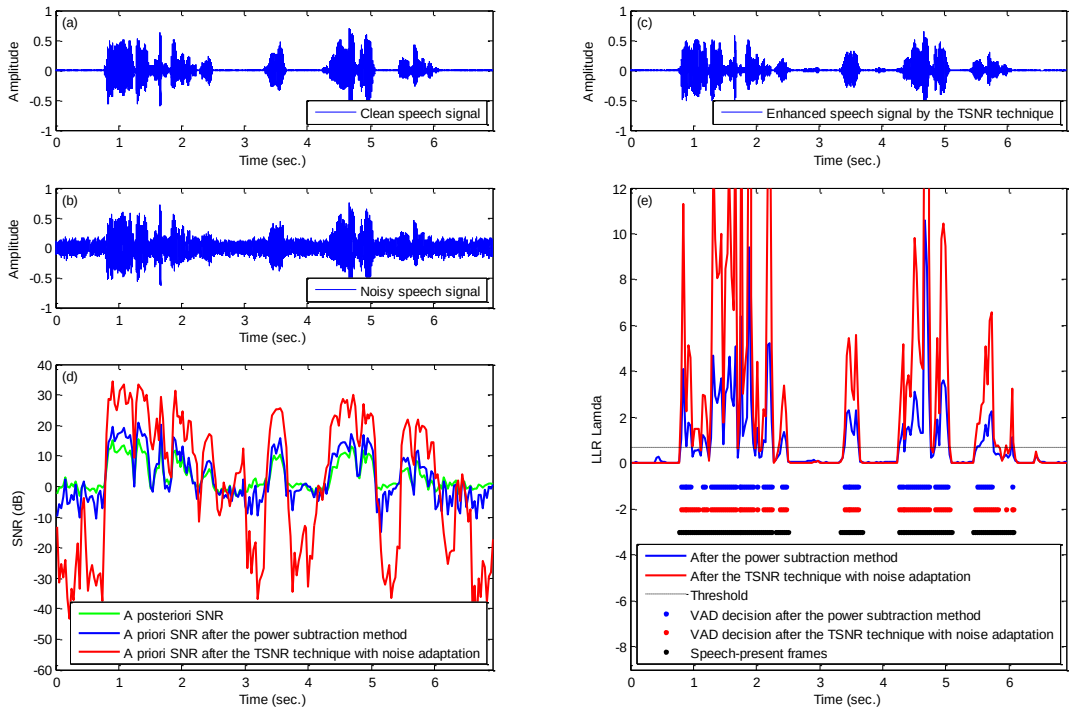

Figure 3.2: Results of VAD using the power subtraction method and the TSNR technique with recursive noise adaptation: (a) clean speech signal, (b) noisy signal with music, (c) enhanced speech signal by the TSNR technique, (d) a posteriori SNR and *a priori* SNRs, (e) LLR Λ and VAD decisions.

system of the SIG-2 humanoid robot.

3.4.1 Experiments

An objective test was conducted to evaluate the performance of the proposed statistical model-based VAD algorithm. The results are shown in Fig. 3.2. Speech signals of a male and a female who speak Japanese were recorded for 7 seconds (Fig. 3.2-(a)) with a 16 kHz sampling frequency and then mixed with background music as additive noise (Fig. 3.2-(b)). The sentences of the speech signals, which were used in the conversation, are “kono uwagiwa Suzuki-san nodesuka?” (female) and “ie, sono uwagiwa Lee-san nodesu.” (male).

3.4.2 Experimental Results

The speech signal enhanced by (3.13) in the TSNR technique is shown in Fig. 3.2-(c). The *a posteriori* SNR and the *a priori* SNRs estimated by the power subtraction method in (3.6) and the TSNR technique in (3.15) are shown in green, blue, and red, respectively (Fig. 3.2-(d)). The VAD decisions by (3.8) in the existing statistical model-based VAD algorithm and (3.16) in the proposed VAD algorithm are also shown in blue and red (Fig. 3.2-(e)).

As the results of the test shows, the proposed VAD algorithm improves *a priori* SNRs by 5.31 dB from those of the existing statistical model-based VAD algorithm on the speech-present period. These improved *a priori* SNRs make LLR values for the VAD decision more affluent and it reduces the detection errors of speech-present period as the final outcome. This means that the *a priori* SNR estimation is a key function in improving the performance of the existing statistical model-based VAD algorithm. The proposed statistical model-based VAD algorithm employing the TSNR technique with recursive noise adaptation could distinguish the speech-present and speech-absent frames with 12.28 points higher accuracy (83.33% vs. 71.05%) comparing to the existing statistical model-based VAD process.

3.5 Summary

In this chapter, an improved statistical model-based VAD algorithm employing the TSNR technique with recursive noise adaptation was presented. The performance problem on the existing statistical model-based VAD algorithm is the insufficient *a priori* SNR estimation by the power subtraction method. To obtain a better performance from the existing statistical model-based VAD algorithm, the *a priori* SNR estimation was improved by utilizing the TSNR technique with recursive noise adaptation instead of the power subtraction method. As a result, *A priori* SNR could be improved with the TSNR technique by 5.31 dB on average during speech-present period. Experimental results demonstrated that the proposed statistical model-based VAD algorithm can indicate the presence and absence of speech with 12.28 points higher accuracy (83.33% vs. 71.05%) than the existing statistical model-based VAD algorithm.

CHAPTER 4
Robust Sound Localization for Binaural Robot Audition

4.1 Introduction

This chapter presents an improved SSL method based on the GCC-PHAT method for binaural robot audition. Effective SSL is a key to understanding the acoustic scene and achieving more natural HRI. Many robot audition systems have been developed using the GCC-PHAT method, and their performance has gradually improved. However, most of these robot audition systems utilize their microphone array, which consists of lots of microphones, to protect the localization performance from various technical problems. Since the binaural robot audition system consists of only two microphones embedded in the robot head, there are difficulties in obtaining a performance as good as that when using the microphone array. In this chapter, two problems affecting the localization accuracy with the conventional SSL based on the GCC-PHAT method in binaural robot audition were addressed:

- **Low-resolution TDOA estimation in time domain**: since the GCC-PHAT method estimates TDOA in the time domain as the sample delay, the estimated TDOA is restricted to an integer value of the sample delay. This restriction makes TDOA estimation inaccurate or impossible in some cases.
- **Diffraction of sound waves with multipath interference caused by contours of robot head**: sound waves easily bend around the robot head, resulting in a difference in TDOA between the waves that travel around the front of the head and those that travel around the back of the head

These two problems severely degrade localization performance, especially for sound

sources in the lateral direction (around ±90°). In this chapter, solutions to these two problems that have improved localization performance in binaural robot audition were as follows:

- **ML estimation in frequency domain**: assuming that frequency bins of the cross-power spectrum are obtained from the exponential distribution with the characteristic function, the ML estimation was applied to the GCC-PHAT method in the frequency domain.
- **New time delay factor**: assuming that the robot head is spherical, a new time delay factor was incorporated into the GCC-PHAT method to compensate for the diffraction of sound waves with multipath interference.

These two solutions are implemented and evaluated experimentally in the binaural SSL system of the SIG-2 humanoid robot using the HARK open-sourced robot audition software. Experiments conducted on the SIG-2 humanoid robot show that the improved SSL method with these two solutions outperforms the conventional SSL method; it reduces localization errors by 17.92° on average and by over 35° in the side directions.

The outline of the chapter is as follows: Section 4.2 summarizes the acoustic model of signals reaching the microphones and the conventional SSL based on the GCC-PHAT method. Section 4.3 defines two problems affecting the localization accuracy in binaural robot audition. Section 4.4 presents their solutions to the two problems. Sections 4.5 and 4.6 outline the binaural SSL system and present experimental results with discussions.

4.2 Conventional Sound Source Localization

In this summary of conventional SSL based on the GCC-PHAT method, an *F*-point STFT with a far-field assumption was used. It was considered the frequency and phase content of local sections of sound signals to change over time and assumed that the acoustic signals reaching each microphone have parallel incidence for SSL. Since acoustic signals consist of varied changes in pitch, volume, timber, and tone over time and since they usually occur far from the microphones in a localization situation, the STFT and the far-field assumption have been generally used in SSL.

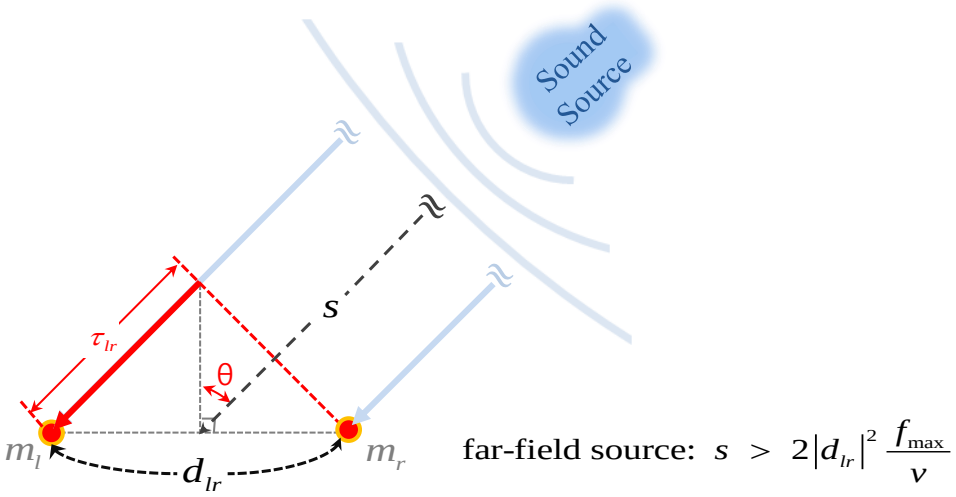

Figure 4.1: Time delay estimation in free space environment.

4.2.1 Acoustic Model

In an ideal scenario, the signals reaching the left and right microphones in a situation with a single sound source can be mathematically modeled as

$$\begin{aligned} X_l[f,n] &= \alpha_l[f]\,|S[f,n]|\exp\left(-j2\pi\frac{f}{F}fs\tau_l\right)+N_l[f,n] \\ X_r[f,n] &= \alpha_r[f]\,|S[f,n]|\exp\left(-j2\pi\frac{f}{F}fs\tau_r\right)+N_r[f,n], \end{aligned} \tag{4.1}$$

where $X_{l,r}[f,n]$, $S[f,n]$, and $N_{l,r}[f,n]$ are the f-th elements of the STFT of the measured signals reaching the two microphones (l and r), the sound source, and uncorrelated additive noise, respectively, on the n-th time frame index. The $f \in \{1, \ldots, F\}$ denotes a frequency bin, F is the frame size of the STFT , fs is the sampling frequency; $\alpha_{l,r}$ and $\tau_{l,r}$ are the attenuation factor and time delay from the position of the sound source to each microphone, respectively. The TDOA between the two microphones is geometrically defined by the relationship in (4.1) and by the free space environment as shown in Fig. 4.1, using microphone l as a reference under the far-field assumption:

$$\tau_{lr} = \tau_r - \tau_l = \frac{d_{lr}}{v}\sin\left(\frac{\theta}{180}\pi\right), \tag{4.2}$$

where τ_{lr} denotes the difference between time delays τ_l and τ_r, d_{lr} is the distance between the two microphones, $\theta \in \{-90°, \ldots, +90°\}$ is the direction of sound incidence in which we are interested, and v is the speed of sound (340.5 m/s at 15℃ in air).

4.2.2 Generalized Cross-Correlation Method with Phase Transform Weighting

The direction of sound incidence in SSL is obtained by first estimating TDOA. A commonly used method for estimating τ_{lr} from unknown parameters τ_l and τ_r is the GCC-PHAT method, which is defined as

$$\hat{R}_{x_l x_r}[\tau_{lr}, n] = \sum_{f=1}^{F} G^{PHAT} X_l[f, n] X_r^*[f, n] \exp\left(j2\pi \frac{f}{F} fs\tau_{lr} \right), \tag{4.3}$$

where

$$G^{PHAT} = \frac{1}{\left| X_l[f, n] X_r^*[f, n] \right|}, \tag{4.4}$$

$\hat{R}x_l x_r$ is the estimate of the cross-correlation function, * represents the complex conjugate, and G^{PHAT} is a normalization factor that preserves only the phase information. The cross-correlation function is calculated as the inverse Fourier transform of the cross-power spectrum with PHAT weighting used for its computational efficiency [86]:

$$csp_{lr}[t, n] = ISTFT(G^{PHAT} X_l[f, n] X_r^*[f, n]), \tag{4.5}$$

where csp_{lr} is the coefficient of the CSP analysis, t is the index of time samples, and $ISTFT$ is the inverse STFT. As the coefficient of the CSP analysis represents a delta pulse centered on the delay, τ_{lr} is estimated as

$$\hat{\tau}_{lr}[n] = \arg\max_t (csp_{lr}[t, n]) \frac{1}{fs}. \tag{4.6}$$

After τ_{lr} is obtained, the direction of sound incidence is estimated using (4.2), which is rewritten as

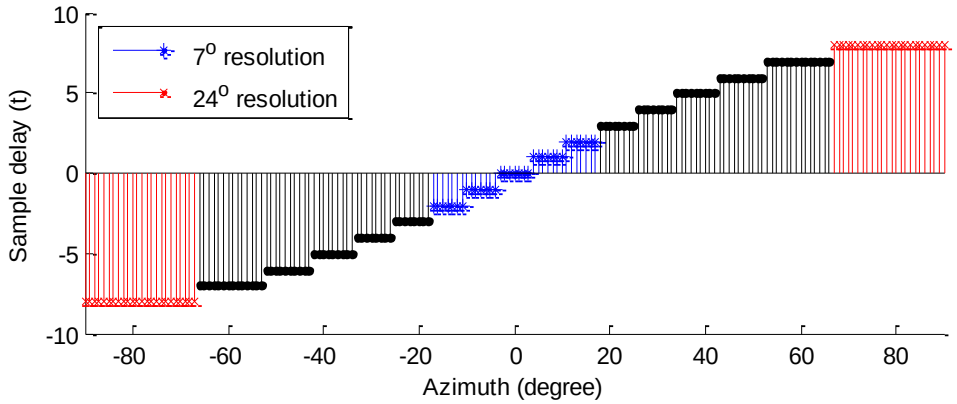

Figure 4.2: Sample delays corresponding to TDOAs along azimuth from −90° to +90°.

$$\hat{\theta}[n] = \sin^{-1}\left(\hat{\tau}_{lr}[n]\frac{v}{d_{lr}}\right)\frac{180}{\pi}. \tag{4.7}$$

4.3 Two Problems Affecting Localization Accuracy

The two problems related to the accuracy of the conventional SSL in binaural robot audition are explained here in detail.

4.3.1 Low-Resolution TDOA Estimation in Time Domain

In conventional SSL, the TDOA is estimated as the sample delay in the time domain using CSP analysis as formulated in (4.5). This restricts the estimated TDOA to an integer value of the sample delay, meaning that SSL is inaccurate or even impossible in some cases. For instance, if the sampling frequency is 16 kHz, the resolution of the sample delay used to estimate TDOA is limited to 62.5 μs (1 s / 16 kHz). Since the difference in TDOA between signals coming from 80° and 90° is less than 62.5 μs when the two microphones are at least 35 cm apart (so that both TDOAs are estimated from the sample delay formulated in (4.6)), the system cannot distinguish the two sound sources.

Figure 4.2 shows these restrictions on sample delay in fixed-point arithmetic,

corresponding to each TDOA along the azimuth from −90° to +90°. This chart was simulated with two microphones placed 17.4 cm apart in the binaural audition system of the SIG-2 humanoid robot, where the far-field assumption holds for any sound source at a distance greater than 1.43 m with a sampling frequency of 16 kHz [87]. As Fig. 1 shows, the best resolution with which the system can distinguish the directions of sound incidence is the 7° resolution in the central direction and the worst is 24° resolution in the lateral directions (−90° to −67 ° and +67° to +90°).

Simple solutions to this problem are to widen the distance between the two microphones, to increase the sampling frequency, or to add extra microphones. However, these solutions have inherent limitations in binaural robot audition with fixed head shape and processing power.

4.3.2 Multipath Interference due to Diffraction of Sound Waves caused by Shape of Robot Head

Basically, the TDOAs are estimated under the assumption that the microphones are located in free space. However, this assumption is not applicable to TDOA estimation using two microphones in a robot head because the sound waves easily bend and spread along the contours of the robot head, which creates sound diffraction and a difference in TDOA between the waves that travel around the front of the head and those that travel around the back of the head. Figure 4.3 illustrates the two paths created by the diffraction of the sound waves with the assumption that the robot head is spherical. It clearly shows that these two diffracted sound-wave paths and multipath interference must be considered if more accurate SSL in binaural robot audition is to be attained.

4.4 Improved Sound Source Localization by Solving Two Problems

The novel solutions to the two problems with conventional SSL described above are presented here. In binaural SSL, the two problems cause inaccurate and unreliable localizations, especially for sound sources in a lateral direction (around ±90°). This is because the localization errors created by the two problems increase as the direction of

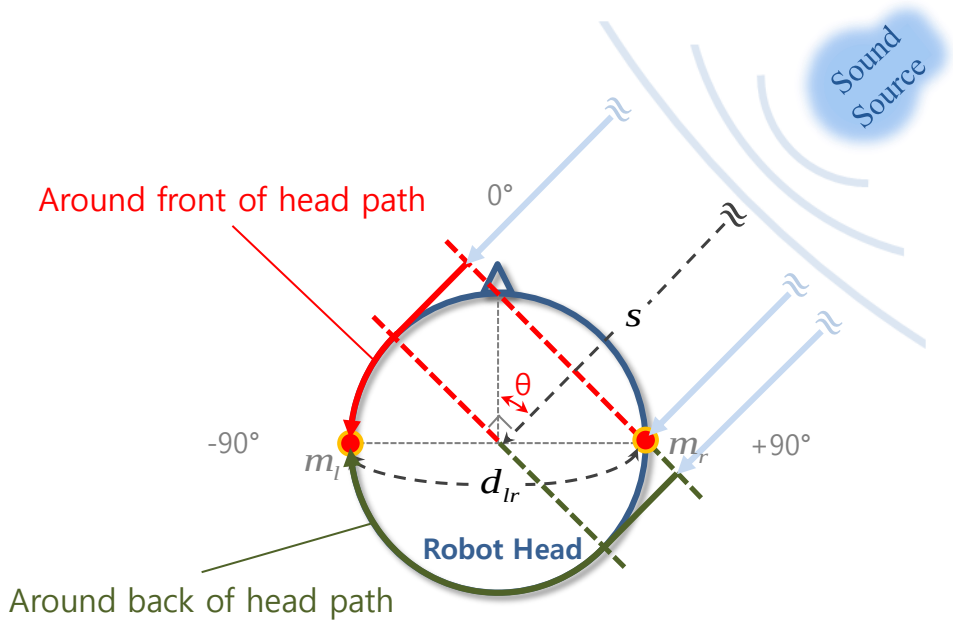

Figure 4.3: Multipath interference due to diffraction of sound waves with spherical-head assumption.

sound incidence approaches −90° or +90°. For accurate binaural SSL, the two problems must be solved.

4.4.1 ML-based SSL in Frequency Domain

To solve the problem of low-resolution TDOA estimation in the time domain, the ML estimation was applied to the conventional SSL in the frequency domain. Assuming that frequency bins f_1, f_2, …, f_T of F, obtained from the exponential distribution with the characteristic function in (4.5), are independent and identically distributed (i.i.d.) observations, and the parameter of the distribution θ, referred to as the true sound incidence direction, is the unknown parameter, an ML-based SSL method can be derived by integrating (4.2)–(4.7) without the inverse Fourier transform analysis:

$$\hat{\theta}_{mle}[n] = \arg\max_{\theta} \frac{1}{F} \sum_{f=1}^{F} \frac{X_l[f,n]X_r^*[f,n]}{\left|X_l[f,n]X_r^*[f,n]\right|} \exp\left(j2\pi \frac{f}{F} fs \frac{d_{lr}}{v} \sin\left(\frac{\theta}{180}\pi \right) \right), \qquad (4.8)$$

where the estimated sound incidence direction θ_{mle} can be obtained by finding the θ that

maximizes the sum of the cross-power spectrum with PHAT weighting in the frequency domain. This ML-based SSL method overcomes the problem of low-resolution TDOA estimation in the time domain and has two advantages contributing to improved accuracy: 1) selective resolution of SSL by adjusting the interval of the sound incidence directions and 2) individual calculations on frequency bands by summing only the frequency bins of interest in the cross-power spectrum, e.g., summing the frequency bins from around 60 to 7000 Hz for the human voice.

4.4.2 New Time Delay Factor

To solve the diffraction with multipath interference problem, simplified formulas to the two paths were first applied under the assumption that the head is spherical:

$$\tau_{front}(\theta) = \frac{d_{lr}}{2v}\left(\frac{\theta}{180}\pi + \sin\left(\frac{\theta}{180}\pi\right)\right), \tag{4.9}$$

$$\tau_{back}(\theta) = \frac{d_{lr}}{2v}\left(\mathrm{sgn}(\theta)\pi - \frac{\theta}{180}\pi + \sin\left(\frac{\theta}{180}\pi\right)\right), \tag{4.10}$$

where τ_{front} and τ_{back} are respectively the time delay for the path around the front of the head and that for the one around the back of the head for each sound incidence direction, and *sgn* is a signum function that extracts the sign of θ; i.e., if θ has a negative sign, $sgn(\theta)$ is −1. After formulas for the two paths are derived, the time difference between them for each sound direction is obtained using

$$\tau_{diff}(\theta) = \tau_{back}(\theta) - \tau_{front}(\theta) = \frac{d_{lr}}{2v}\left(\mathrm{sgn}(\theta)\pi - \frac{2\theta}{180}\pi\right), \tag{4.11}$$

where τ_{diff} is 0 when θ is −90° or +90°. Suppose that the intensity of the multipath interference from τ_{back} for each sound direction corresponds to that of the ILD ratios between the two microphones in the robot head, where the ILD ratios represents the sine function in the ideal condition. τ_{diff} multiplied by the absolute sine function with attenuation factor β_{multi} (typically set to 0.1) was used as the factor used to compensate for the multipath interference:

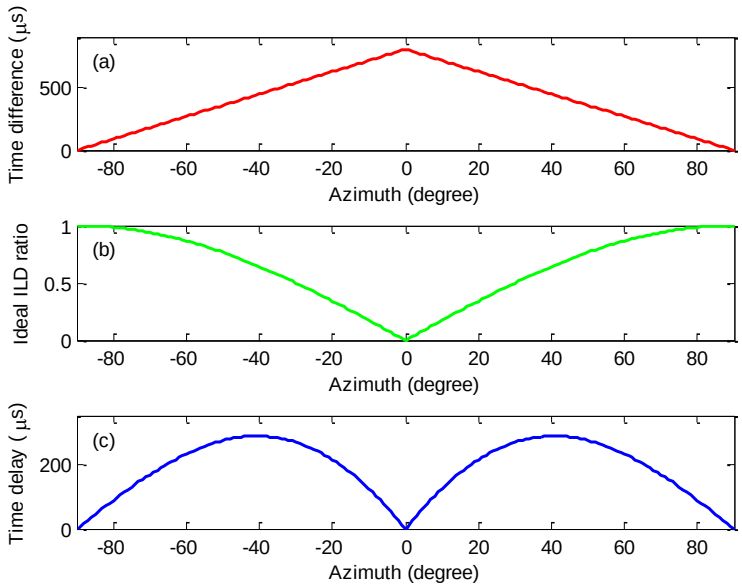

Figure 4.4: Deriving compensation factor for multipath interference: (a) absolute values of different time delays between two paths along front and back heads, (b) absolute values of ILD ratios complied with sine function, (c) time delays calculated from (a) multiplied by (b) to compensate multipath interference.

$$\tau_{\text{inter}}(\theta) = \frac{d_{lr}}{2v}\left(\text{sgn}(\theta)\pi - \frac{2\theta}{180}\pi\right)\cdot\left|\beta_{\text{multi}}\sin\left(\frac{\theta}{180}\pi\right)\right|, \tag{4.12}$$

where τ_{inter} is the interference created by the two paths. This derived τ_{inter} is shown in Fig. 4.4, which was simulated with a pair of microphone 17.4 cm apart. The final time delay factor to be used instead of $\tau_{lr}(\theta)$ in (4.2) for binaural SSL can be derived using τ_{front} and τ_{inter}:

$$\begin{aligned}\tau_{multi}(\theta) &= \tau_{front}(\theta) - \tau_{\text{inter}}(\theta) \\ &= \frac{d_{lr}}{2v}\left(\frac{\theta}{180}\pi + \sin\left(\frac{\theta}{180}\pi\right)\right) - \frac{d_{lr}}{2v}\left(\text{sgn}(\theta)\pi - \frac{2\theta}{180}\pi\right)\cdot\left|\beta_{multi}\sin\left(\frac{\theta}{180}\pi\right)\right|.\end{aligned} \tag{4.13}$$

This new time delay factor, τ_{multi}, is used with the ML-based SSL method in (4.8):

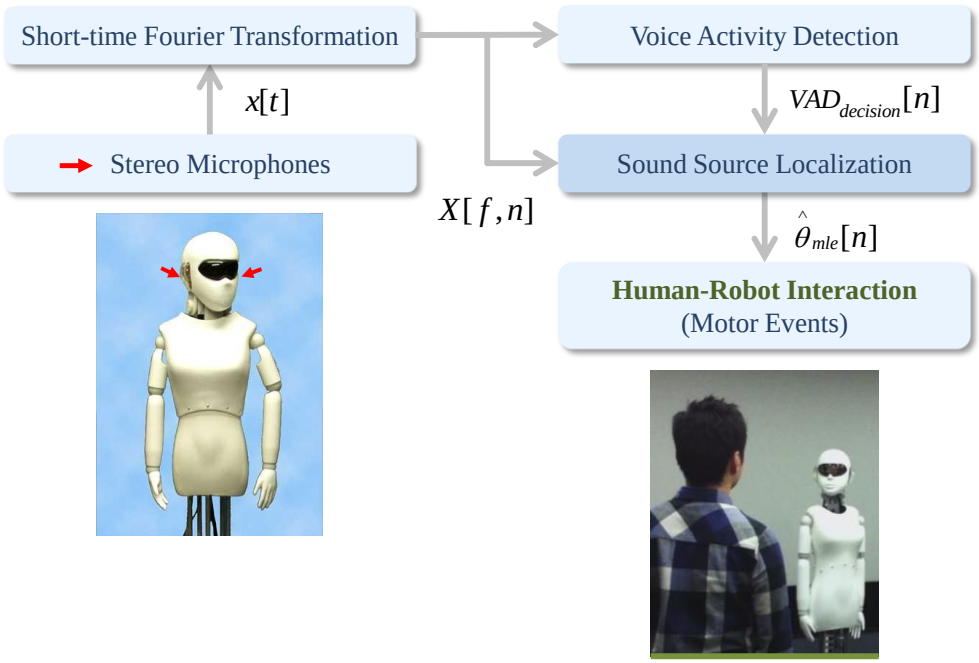

Figure 4.5: Flow of SSL in SIG-2 humanoid robot with improved methods.

$$\hat{\theta}_{mle}[n] = \arg\max_{\theta} \frac{1}{F} \sum_{f=1}^{F} \frac{X_l[f,n]X_r^*[f,n]}{\left|X_l[f,n]X_r^*[f,n]\right|} \exp\left(j2\pi \frac{f}{F} fs\tau_{multi}(\theta) \right). \quad (4.14)$$

4.5 Evaluation

The ML-based SSL method with time delay factor τ_{multi} was evaluated in various ways to verify that it makes fewer localization errors than the conventional SSL method in binaural robot audition. Figure 4.5 shows the flow of this implemented SSL system. For the body of the system, the improved statistical model-based VAD algorithm derived in Chapter 2 was also used a significant building block to deactivate SSL process during speech-absent period to reduce the computational cost and an unexpected SSL error. The estimated sound incidence directions were used to make the robot turn at its neck and waist in order to look in the speaker's directions.

4.5.1 Experiments

The experiments were conducted in a room (6 m long × 4.25 m wide × 2.85 m high) with a reverberation time of about 120 ms and noise from two air conditioners installed on each corner of the room and two personal computers placed on the other corners. The SIG-2 humanoid robot using the Sennheiser ME 104 omnidirectional microphone was placed at the center of the room, and the speakers were located 1.5–2.5 m from the robot. To create a noisier environment, background music with lyrics was played on a laptop just below the robot as additive noise. The average sound pressure level (SPL) of the background music and the average SNR of the target speech signals were about 62.7 dB and 21.2 dB, respectively. The system recorded the background noise for 2 s before each trial to estimate the noise variance and used the variance as the *a priori* noise variance for the VAD process.

To obtain an accurate estimate of the performance improvement with the ML-based SSL method compared to the conventional SSL method in binaural robot audition, it is estimated in various ways with a static or moving sound source. A male and then a female speaker stood at points along the azimuth from −90° to +90° in 10° steps and spoke to the robot five times at each point. Their speech signals were captured using the four experimental methods:

- The conventional SSL method using (4.3)–(4.7) described in Section 4.2.
- The ML-based SSL method using (4.8) as the solution to the first problem described in Section 4.3.1.
- The ML-based SSL method using (4.9) as the solutions to the first and second problems without considering multipath interference in Section 4.3.
- The ML-based SSL method using (4.13) and (4.14) as the solutions to the first and second problems with considering multipath interference described in Section 4.3.

The ML-based SSL method using (4.13) and (4.14) was also evaluated with a moving speaker who had changed average walking speed three times at 0.06 m/s, 0.11 m/s, and 0.20 m/s (the average walking speed of healthy adults is 1.0 m/s) for 250 s to verify its real-time processing and effectiveness. For this evaluation of moving speaker situation, the OptiTrack motion capture system had used as a ground truth tool and had captured the moving path of the speaker with trackballs concurrently [88].

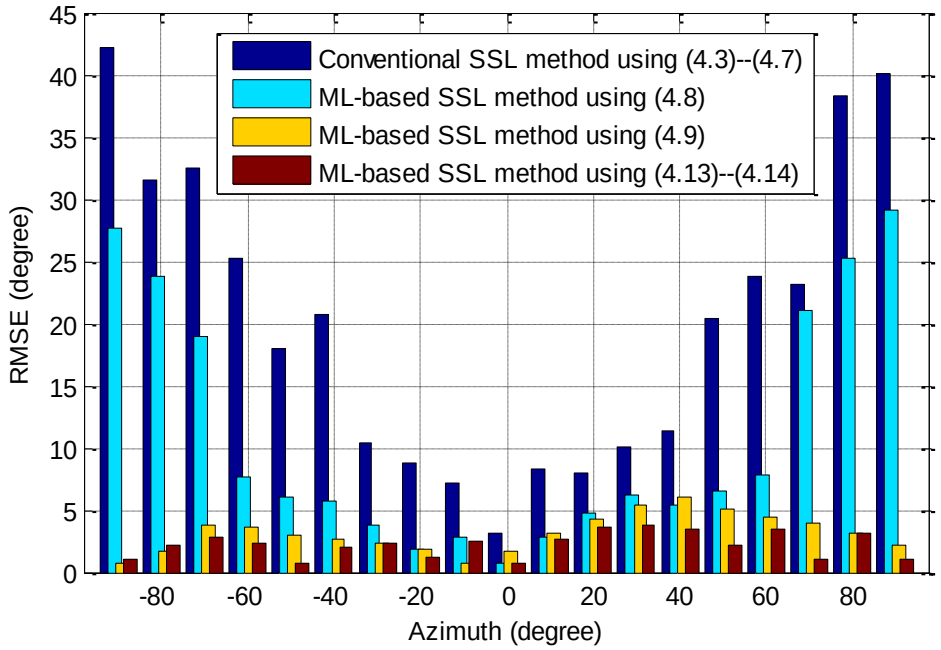

Figure 4.6: RMSEs of 190 trials for four SSL methods.

4.5.2 Experimental Results

Figure 4.6 shows the root-mean-square-error (RMSE) for the 190 trials (19 points × 5 speech signals × 2 speakers) for the four experimental methods. As shown in the figure, the two ML-based SSL methods had fewer localization errors than the conventional method. The one using (4.13) and (4.14) was particularly effective—it reduced the average RMSE by 17.92° (2.23° vs. 20.15°) and the RMSEs for the side directions by over 35°.

The results represented by the cyan bars in Fig. 4.6 demonstrate that applying the ML-based SSL method in the frequency domain improves localization performance. The results represented by the orange bars in Fig. 4.6 demonstrate that the effect of multipath interference along the shape of the robot head in SSL can be identified by observing parabolic increases in RMSEs when the ML-based SSL using (4.9) had been tested. These parabola increases were almost the same of those of the compensation factor derived in Section 4.4.2 (See Figure 4.4-(c)). The results represented by the red

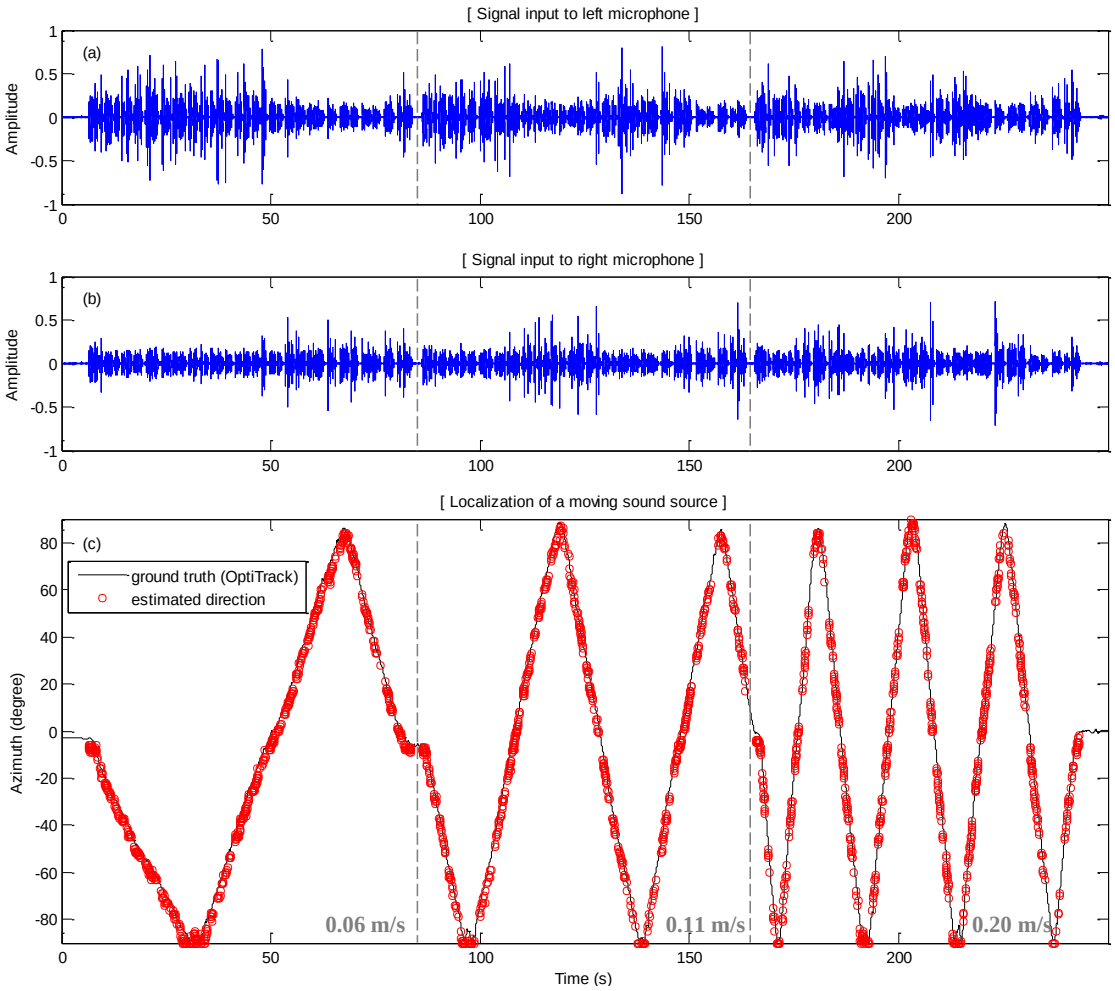

Figure 4.7: Real-time SSL for a moving speaker: (a) (b) signal inputs to left and right microphones consisting of male speech signal; (c) actual speaker directions and estimated directions.

bars in Fig. 4.6 demonstrate that applying the ML-based SSL method using time delay factor τ_{multi} improves localization performance even more, especially for the side directions. This means that localization performance can be significantly improved by considering the diffraction of the sound waves and the multipath interference of the sound signals reaching each microphone in binaural robot audition.

Figure 4.7 shows the moving speaker path changing walking speed three for 250 s, which was measured by OptiTack and the estimated directions by the ML-based SSL method using (4.13) and (4.14). Three average RMSEs for the estimated directions in three different moving speeds at 0.06 m/s, 0.11 m/s, and 0.20 m/s were 1.96º, 2.01º, and

2.54º, respectively.

As a result, despite the proposed SSL system having only two microphones located in the robot head, it showed the overall performance in real time which is as good as other systems utilizing a microphone array consisted of many microphone.

4.6 Summary

Two accuracy problems with conventional SSL based on the GCC-PHAT method in binaural robot audition were addressed: 1) low-resolution TDOA estimation in the time domain and 2) diffraction of sound waves with multipath interference around the robot head. To solve the first problem, the ML estimation in the frequency domain to the GCC-PHAT method was applied instead of CSP analysis which estimates TDOA in the time domain. To solve the second problem, a new time delay factor was developed for use with the GCC-PHAT method that takes into account the diffraction of sound waves and multipath interference under the assumption that the robot head is spherical.

Experimental results demonstrated that the ML-based SSL method in the frequency domain contributes to improved localization resolution and taking the diffraction of sound waves with multipath interference into account when estimating the time delay is a key to improving SSL performance in binaural robot audition.

CHAPTER 5
Binaural Sound Localization over Entire Azimuth

5.1 Introduction

This chapter presents an improved SSL method over the entire azimuth by front-back disambiguation for use with binaural robot audition equipped with two microphones inside artificial human-like pinnae. The problem of front-back ambiguity in binaural robot audition was addressed, which limits the localization range to the front horizontal space from −90° to +90°:

- **Front-back ambiguity due to the same TDOA for the front and back**: binaural audition methods localize a sound source as coming from the front despite the actual sound source being in the rear.

Current methods to solve this ambiguity problem involve using head movements and using a specific HRTF database. However, these methods have certain drawbacks. Using head movement does not work well for short words or phrases because the robot needs enough time to complete its head movement and it causes self-motor noise in motion. Using an HRTF database does not work well if the system and environment change because its performance depends greatly on the system and environment. In this chapter, a different approach to solving this problem was taken for SSL over the entire azimuth:

- **Pinna amplification effect**: the pinna amplification effect, which creates a level difference between sound signals coming from the front and back, was utilized.

This solution was implemented and experimentally evaluated in the binaural audition system of the SIG-2 humanoid robot. Experiments conducted using the SIG-2 humanoid

robot showed that SSL errors over the entire azimuth were reduced by 9.9° on average with the new time delay factor proposed in Chapter 4, compared to using the conventional time delay factor, and that the success rate for the proposed front-back disambiguation method was 32.2% better on average over the entire azimuth than with a conventional HRTF-based method.

The chapter is organized as follows: Section 5.2 describes the problem of front-back ambiguity in binaural robot audition. Section 5.3 describes the solution to the ambiguity problem: front-back disambiguation using the pinna amplification effect. Section 5.4 describes the evaluation experiments and presents the results. Section 5.5 concludes the chapter with a summary.

5.2 Problem of Front-Back Ambiguity

Normally, the direction of a sound source estimated using a microphone array corresponds to the actual direction. However, a binaural audition system has an inherent problem—a sound source appears to be at equal (mirror) angles in the front and rear hemi-fields due to having the same TDOA, as shown in Fig. 5.1. For example, a sound source placed at 30° (where 0° is directly in front) is estimated to also be at -150° in front-back ambiguity. This front-back ambiguity limits the localization range of a binaural audition system to the front horizontal space from −90° to +90°.

Current solutions to this ambiguity problem in binaural robot audition are to use head movements or HRTF databases. For the use of head movements, once the robot head started to turn, the sound source behind the robot got closer to the left microphone, effectively moving towards the left. At the same time, the sound source in front got closer to the right microphone, moving in the opposite direction. The HRTF database, a series of points where ITD and ILD for sound sources from many locations, is involved in resolving the problem of front-back ambiguity because when a sound is received by the microphones it can be reflected off the pinnae and the robot head that causes different spectral modifications between sound sources in the front and back. However, these two methods have certain drawbacks. Using head movements does not work well for short words or phrases, such as when someone calls the robot's name, because the robot needs enough time to complete its head movement. Using an HRTF database is highly dependent on its system configuration and environment, i.e., HRTFs are

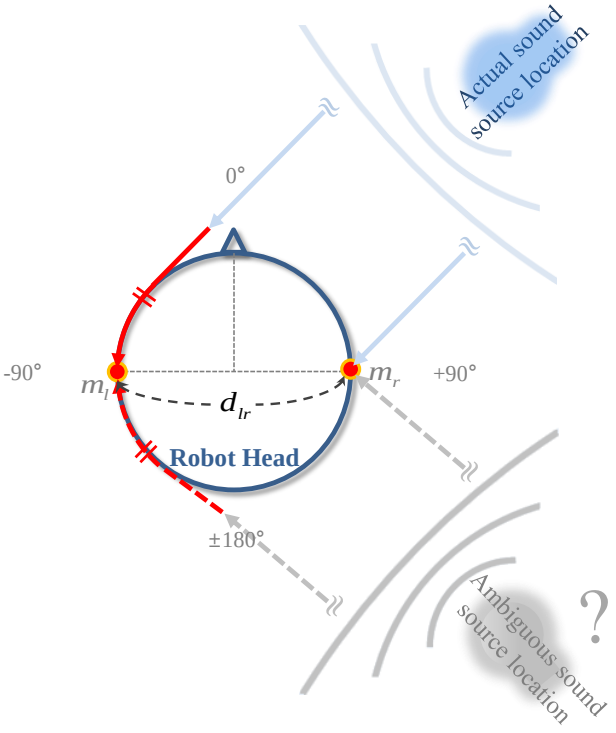

Figure 5.1: Problem of front-back ambiguity in binaural robot audition.

remeasured whenever either the system configuration or environment changes, and HRTFs need to be measured in 1° step for high-resolution localization.

To extend the localization range of binaural robot audition systems over the entire azimuth, the problem of front-back ambiguity needs to be overcome by different approaches.

5.3 Front-Back Disambiguation by using Amplification Effect of Pinnae

The basic function of the pinnae is to collect sound and spectrally transform it, which enable various types of audio signal processing. They collect sound through a filtering process and frequency-dependent amplification. This amplification increases the sound level by about 10 to 15 dB in the 1.5 to 7 kHz range [89].

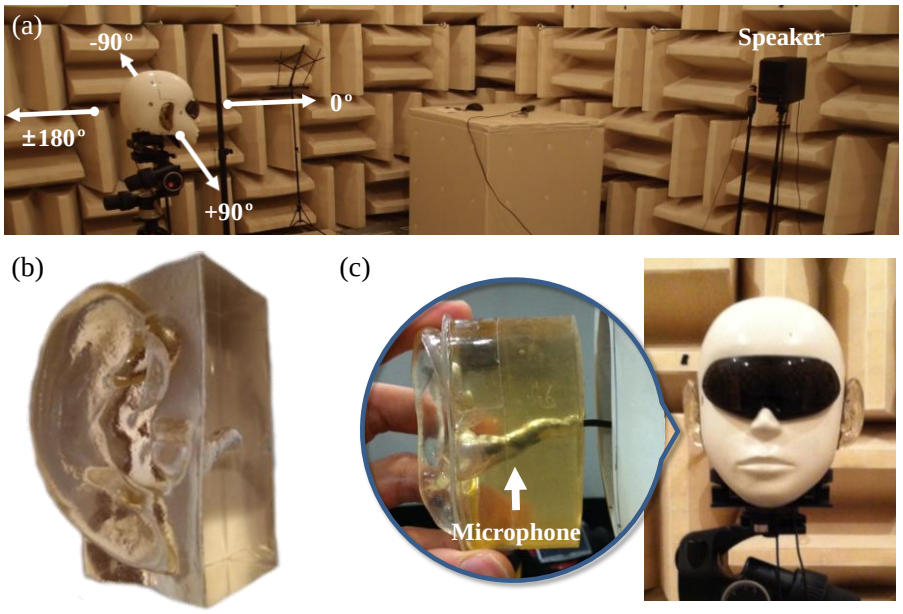

Figure 5.2: HRTF measurement environment with the SIG-2 robot head equipped with pinnae in anechoic chamber: (a) coordinate system for SSL, (b) silicone artificial pinna (7.2 cm long × 3 cm wide × 2.3 cm high), (c) microphone location inside pinna.

The new approach to the front-back ambiguity problem is to utilize this amplification effect of the pinnae. To evaluate sound amplification in the frequency range over the entire azimuth, HRTFs were measured by equipping the head of the SIG-2 humanoid robot with two artificial pinnae and placing it in an anechoic chamber. Figure 5.2 shows the measurement environment, the shape of the pinnae, and the location of the microphone in each pinna. A 0.3-s time stretched pulse (TSP) signal [90] ranging from 1 to 8 kHz was used as a sound source and HRTFs were measured in 5° steps over the entire azimuth. The measured HRTF data are illustrated in Fig. 5.3. The pinnae amplified the intensities of the input sound signal in the frequency band ranging from 3000 to 5300 Hz for the front directions (−90° to +90°). The left pinna amplified it from around −90° to +10° (Fig. 5.3-(a)), and the right one amplified it from around −10° to +90° (Fig. 5.3-(b)). Since this pinna effect amplifies the sound signals coming from only the front directions, the problem of front-back ambiguity can be solved by simply comparing the mean intensity of the sound signals in a specific frequency range with a

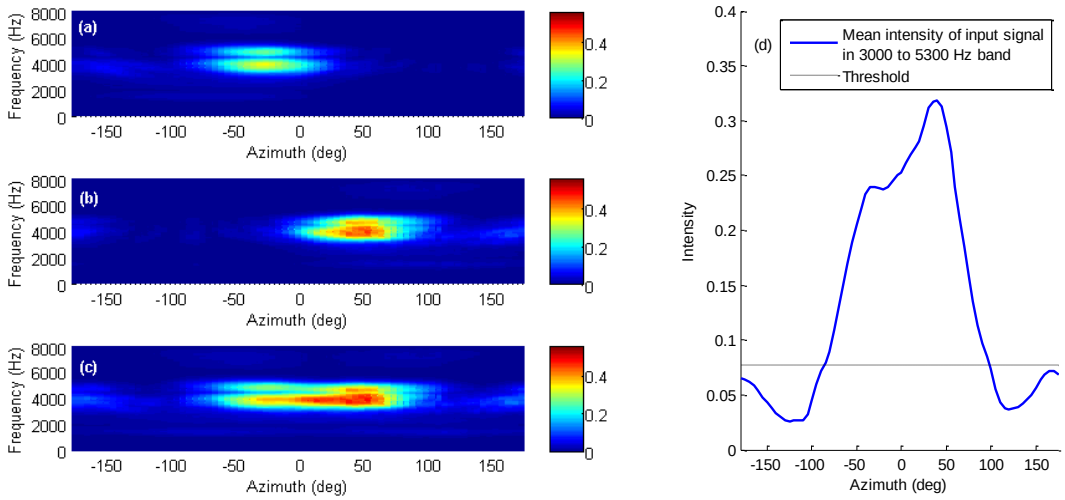

Figure 5.3: Effects of amplification by silicone artificial pinnae: (a) amplification by left pinna, (b) amplification by right pinna, (c) amplification by left and right pinnae, (d) mean intensity of input time stretched pulse signals in 3000 to 5300 Hz frequency band.

threshold value, as shown in Fig. 5.3-(d). The decision rule for front-back disambiguation using the pinna amplification effect is

$$\text{if} \quad \log \frac{1}{f_2^{FB} - f_1^{FB}} \sum_{f=f_1^{FB}}^{f_2^{FB}} \left|X_l[f,n]\right|^2 + \left|X_r[f,n]\right|^2 \; > \; \eta_{FB} \quad \text{then} \quad \text{frontal direction} \qquad (5.1)$$

$$\text{else} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{rear direction} \quad ,$$

where $f_1^{FB}=3000/F_s$ and $f_2^{FB}=5300/F_s$ are the boundary frequency bins to be calculated. The input sound signals are normalized by the mean value of the intensities of all frequency bins beforehand to make the intensities consistent.

If the observed signal is determined to be behind the head by using this decision rule, the direction estimated by the SSL method is switched to the mirrored angle location in the back (e.g., +30° is switched to +150°).

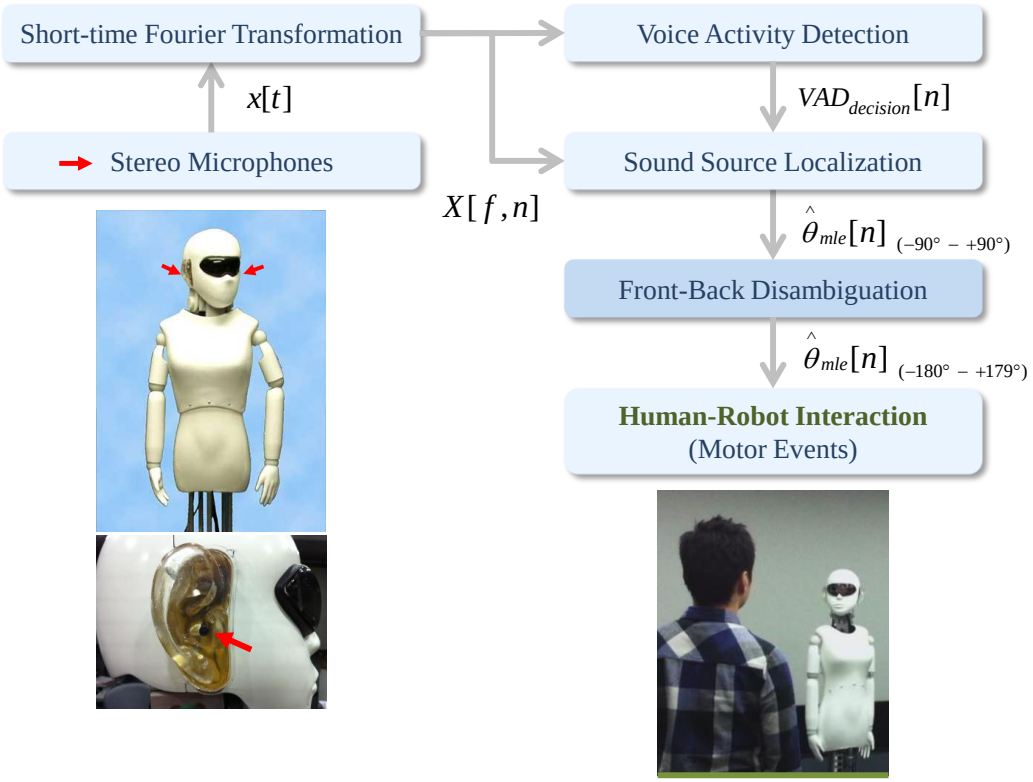

Figure 5.4: Front-back disambiguation by using pinna amplification effect in SIG-2 humanoid robot.

5.4 Evaluation

The front-back disambiguation using pinna amplification effect in (5.1) was evaluated with the improved ML-based SSL method described in Chapter 4 under various conditions to determine whether it enables more effective SSL with front-back disambiguation over the entire azimuth than the one using a conventional HRTF-based method. The proposed disambiguation method was implemented in the binaural audition system of the SIG-2 humanoid robot. Figure 5.4 shows the flow of the SSL process with front-back disambiguation over the entire azimuth. Since the target sound signals to be recognized by robots are usually human speech, localization processing was restricted to only when a human speech was detected by using the improved statistical model-based VAD algorithm derived in Chapter 2.

5.4.1 Experiments

The experiments were conducted in a room (6 m long × 4.25 m wide × 2.85 m high) with a reverberation time of about 120 ms and noise from air conditioners and personal computers. To make the environment noisier, background music with lyrics was played as additive noise. The average sound pressure level (SPL) of the background music and the average SNR of the target speech signals were about 61.2 and 23.2 dB, respectively. The SIG-2 humanoid robot was placed at the center of the room and the speakers were positioned 1.5–2.5 m from the robot in various directions. The system first recorded the background noise for 2 s to estimate the noise variance and used this variance as the *a priori* noise variance for the VAD process. A male and then a female speaker stood at points along the azimuth from −180° to +170° in 10° steps and spoke short words or phrases (for about 1 s; e.g., calling the robot's name and expressing simple greetings) to the robot five times at each point, and the system captured their speech signals.

To accurately estimate the improvement in SSL performance with front-back disambiguation over the entire azimuth, five experiments were conducted using five different methods:

- The ML-based SSL method using conventional time delay factor τ_{lr} in (4.8).
- The ML-based SSL method using derived time delay factor τ_{front} in (4.9) considering only front path around robot head.
- The ML-based SSL method using proposed time delay factor τ_{multi} in (4.13) and (4.14) considering both front and back paths around robot head.
- Front-back disambiguation using a conventional HRTF-based method.
- Front-back disambiguation using proposed method in (5.1).

The HRTF-based method used in the fourth experiment uses a decision rule in which two measures of the cross-correlation coefficient between the input signal and the HRTF data measured in the coincident front-direction or in the mirrored back-direction are compared:

$$\begin{aligned} &\textit{if} \quad R_{XH}^{front}[n] >= R_{XH}^{back}[n] \quad \textit{then} \quad \textit{front direction} \\ &\textit{else} \qquad\qquad\qquad\qquad\qquad\quad \textit{back direction,} \end{aligned} \tag{5.2}$$

where

$$R_{XH}^{front}[n] = \frac{1}{F}\sum_{f=1}^{F} corr\left(\log\left(\frac{|X_r[f,n]|^2}{|X_l[f,n]|^2}\right), \log\left(\frac{|H_r[f,\theta_{front}]|^2}{|H_l[f,\theta_{front}]|^2}\right)\right)$$
$$R_{XH}^{back}[n] = \frac{1}{F}\sum_{f=1}^{F} corr\left(\log\left(\frac{|X_r[f,n]|^2}{|X_l[f,n]|^2}\right), \log\left(\frac{|H_r[f,\theta_{back}]|^2}{|H_l[f,\theta_{back}]|^2}\right)\right), \quad (5.3)$$

$H_{l/r}$ is the HRTF data measured from each of the two microphones for the $\theta_{front/back}$ direction beforehand.

5.4.2 Experimental Results

Figure 5.5 shows the RMSE for the 360 trials (36 points × 5 speech signals × 2 speakers) for the three SSL methods along the azimuth from −180° to +170° in 10° steps. The SSL method using the new time delay factor proposed in Chapter 4 had fewer localization errors over the entire azimuth than the one using time delay factor τ_{lr} derived in free space and the one using time delay factor τ_{front} derived considering only sound wave diffraction. The average RMSEs for τ_{lr}, τ_{front}, and τ_{multi} were respectively 11.87°, 3.40°, and 1.96°.

Figure 5.6 shows the rate of successful disambiguation for the 360 trials for the two disambiguation methods along the entire azimuth. The proposed disambiguation method using the pinna amplification effect had an average success rate with 32.2% higher (92.28% vs. 69.78%) than the one using the HRTF database.

As shown by the red bars in Fig. 5.5, considering the diffraction of sound wave with multipath interference around the robot head effectively improves SSL performance over the entire azimuth in binaural robot audition compared to other SSL methods (blue and green bars). The disambiguation method using the pinna amplification effect proposed in this chapter was better able to disambiguate the directions of the input sound signals coming from either the front or back, as shown in Fig. 5.6. The HRTF-based method had worse disambiguation performance, especially for signals coming from the back, because the measurement of the cross-correlation coefficient using the front-HRTF data was often slightly higher than or the same as that using the back-HRTF data even though the sound signals came from the back. This means that there was not much difference in the frequency properties and spectral cues of the front and back HRTF data for some directions.

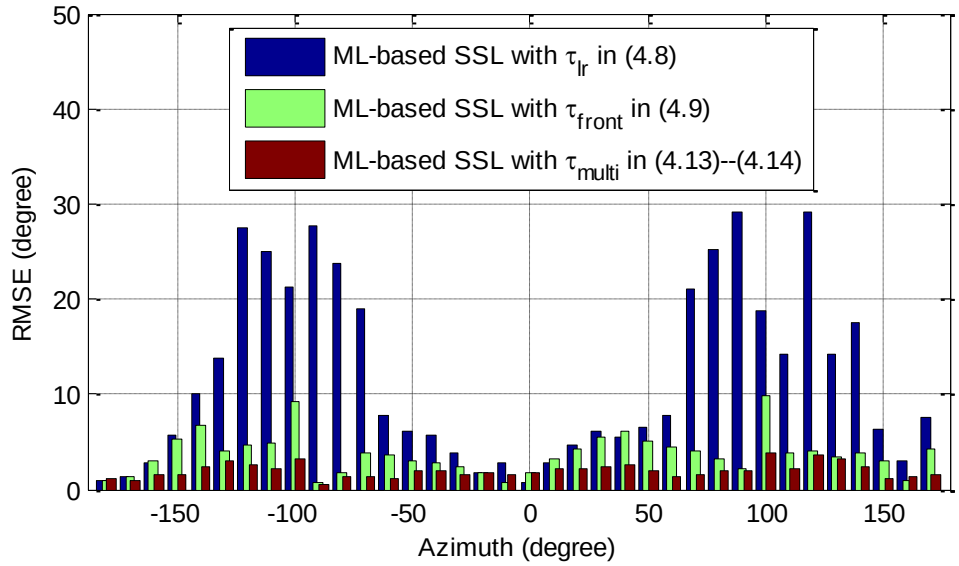

Figure 5.5: RMSEs for 360 trials for three SSL methods over entire azimuth.

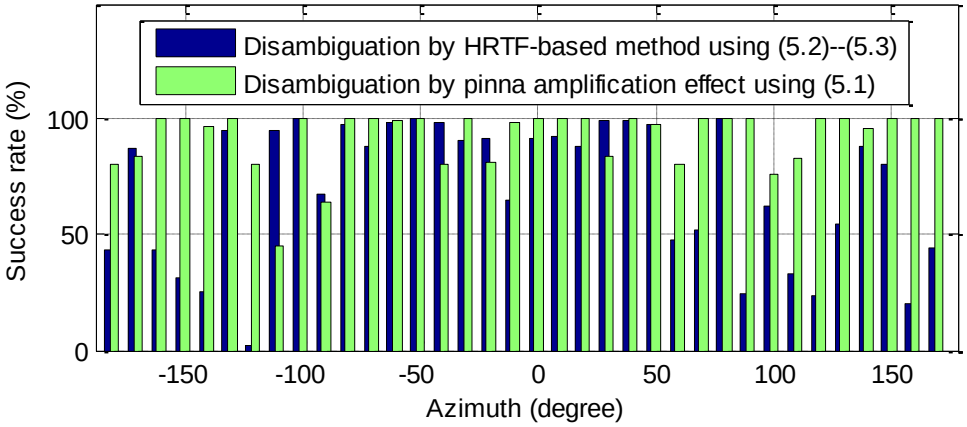

Figure 5.6: Success rates for two disambiguation methods.

5.5 Summary

In this chapter, the problem of front-back ambiguity in binaural robot audition, which limits the localization range to the front horizontal space, was addressed. To solve the problem, a new front-back disambiguation method utilizing the pinna amplification

effect was devised. Experimental results demonstrated that utilizing the pinna amplification effect effectively solves the problem of front-back ambiguity for robots equipped with silicon human-like pinnae. As a result, the implementation of the proposed front-back disambiguation method using the pinna amplification effect with the ML-based SSL method proposed in Chapter 4 enabled accurate binaural SSL over the entire azimuth. The video demonstration is available on Youtube [91].

CHAPTER 6
Binaural Localization of Multiple Sound Sources

6.1 Introduction

This chapter presents a multisource sound localization method based on GCC-PHAT method for binaural robot audition. Another significant problem to be overcome for binaural robot audition was addressed here:

- **Difficulty with multisource sound localization**: since localization performance generally drops as the number of microphones is reduced and binaural robot audition uses only two microphones embedded on each side of the robot head and, the number of sound sources that the binaural system can localize has been limited to a single source in real environments, including reverberations and noises.

The most commonly used multisource sound localization method, MUSIC, is unable to localize multiple sources with binaural sound inputs. In addition, most binaural methods based on their specific HRTFs for multisource sound localization are still with the problem of limited localization resolution and performance in real environments.

In this chapter, the solution to realizing multisource sound localization for binaural robot audition is threefold:

- **Multisource sound localization**: the ML-based SSL method was extended to enable simultaneous multidirection estimations for each time frame.
- **SNR-weighting function**: SNR-weighting function was incorporated into the extended ML-based SSL method to cope with additive noise.
- **Improved *K*-means clustering**: *K*-means clustering was performed in order to eliminate incorrect SSL errors and to estimate an unknown time-varying number of

sound sources.

The ML-based SSL method proposed in Chapter 4 was extended to enable simultaneous multiple direction estimations with the problem of correctly estimating the sound incidence directions and the unknown time-varying number of sound sources in real environments. The degraded SSL performance when there are multiple sound sources was improved by using a SNR-weighting function and an improved *K*-means clustering. For effective multisource sound localization, the standard *K*-means clustering algorithm was improved by adding two additional steps that increase the number of clusters automatically and eliminate clusters that contain incorrect direction estimations.

The proposed multisource sound localization method was implemented as a real-time system using the HARK open-sourced robot audition software and evaluated experimentally in the binaural audition system of the SIG-2 humanoid robot. Experiments conducted in real environments showed that the proposed method can localize multiple sound sources in real time with localization error below 5.96°

The chapter is outlined as follows: Section 6.2 describes the extension of the ML-based SSL method to a multisource situation. Section 6.3 addresses its difficulties with estimating correct multiple directions in real environments. Section 6.4 gives the solution to the difficulties with multisource sound localization: an SNR-weighting function and an improved *K*-means clustering algorithm. Section 6.5 presents the experimental results. Section 6.6 concludes the chapter.

6.2 Extension of ML-Based SSL method to multiple sound sources

The observed signals from the left and right microphones in a situation with *K* sound sources can again be mathematically modeled from (4.1) as

$$\begin{aligned} X_l[f,n] &= \sum_{k=1}^{K} \alpha_{li}[f] \left| S_k[f,n] \right| \exp\left(-j2\pi \frac{f}{F} \tau_{lk} \right) + N_l[f,n] \\ X_r[f,n] &= \sum_{k=1}^{K} \alpha_{ri}[f] \left| S_k[f,n] \right| \exp\left(-j2\pi \frac{f}{F} \tau_{rk} \right) + N_r[f,n], \end{aligned} \tag{6.1}$$

where *k* denotes the index of each sound source. For the ideal case, the multiple sound sources S_k are uncorrelated with each other and with additive noise $N_{l,r}$; i.e.,

$E\{S_1[f,n]S_2[f,n]\}=0$, $E\{S_k[f,n]N_{l,r}[f,n]\}=0$, and $E\{N_l[f,n]N_r[f,n]\}=0$, where $E[.]$ is the expectation operator for time frames. Accordingly, the cross-power spectrum and the denominator of PHAT weighting for multiple sound sources are expressed by using Euler's formula and the Pythagorean trigonometric identity as

$$X_l[f,n]X_r^*[f,n]=\sum_{k=1}^{K}\alpha_{lk}[f]\alpha_{rk}[f]\left|S_k[f,n]\right|^2\exp\left(j2\pi\frac{f}{F}fs(\tau_{rk}-\tau_{lk})\right), \tag{6.2}$$

$$\begin{aligned}
&\left|X_l[f,n]X_r^*[f,n]\right| \\
&=\sum_{k=1}^{K}\alpha_{lk}[f]\alpha_{rk}[f]\left|S_k[f,n]\right|^2\left|\cos\left(2\pi\frac{f}{F}fs(\tau_{rk}-\tau_{lk})\right)+j\sin\left(2\pi\frac{f}{F}fs(\tau_{rk}-\tau_{lk})\right)\right| \\
&=\sum_{k=1}^{K}\alpha_{lk}[f]\alpha_{rk}[f]\left|S_k[f,n]\right|^2\sqrt{\cos^2\left(2\pi\frac{f}{F}fs(\tau_{rk}-\tau_{lk})\right)+\sin^2\left(2\pi\frac{f}{F}fs(\tau_{rk}-\tau_{lk})\right)} \\
&=\sum_{k=1}^{K}\alpha_{lk}[f]\alpha_{rk}[f]\left|S_k[f,n]\right|^2,
\end{aligned} \tag{6.3}$$

where $\alpha_{lk}[f]\alpha_{rk}[f]|S_k[f,n]|^2 \geq 0$ and the TDOA of each sound source obeys the relationship $\tau_{lr}=\tau_{rk}-\tau_{lk}$.

After (6.2) and (6.3) are substituted into (4.14), the ML-based SSL method for multiple sound sources can be expressed as

$$\begin{aligned}
&\arg\max_{\theta}\frac{1}{F}\sum_{f=1}^{F}\frac{X_l[f,n]X_r^*[f,n]}{\left|X_l[f,n]X_r^*[f,n]\right|}\exp\left(j2\pi\frac{f}{F}fs\tau_{multi}(\theta)\right) \\
&=\arg\max_{\theta}\frac{1}{F}\sum_{f=1}^{F}\left(\frac{\sum_{k=1}^{K}\alpha_{lk}[f]\alpha_{rk}[f]\left|S_k[f,n]\right|^2\exp\left(j2\pi\frac{f}{F}fs(\tau_{rk}-\tau_{lk})\right)}{\sum_{k=1}^{K}\alpha_{lk}[f]\alpha_{rk}[f]\left|S_k[f,n]\right|^2}\cdot\exp\left(j2\pi\frac{f}{F}fs\tau_{multi}(\theta)\right)\right) \\
&=\arg\max_{\theta}\frac{1}{F}\sum_{f=1}^{F}const.\cdot\sum_{k=1}^{K}\exp\left(j2\pi\frac{f}{F}fs(\tau_{rk}-\tau_{lk})\right)\exp\left(j2\pi\frac{f}{F}fs\tau_{multi}(\theta)\right),
\end{aligned} \tag{6.4}$$

where TDOAs ($\tau_{rk}-\tau_{lk}$) of K sound sources exist independently and the K directions corresponding to them are represented as a set of K maximum values. Thus, multiple directions θ_{mle_k} for the different TDOAs can be estimated by finding each expected sound incidence direction that makes each peak in the response of the ML-based SSL method

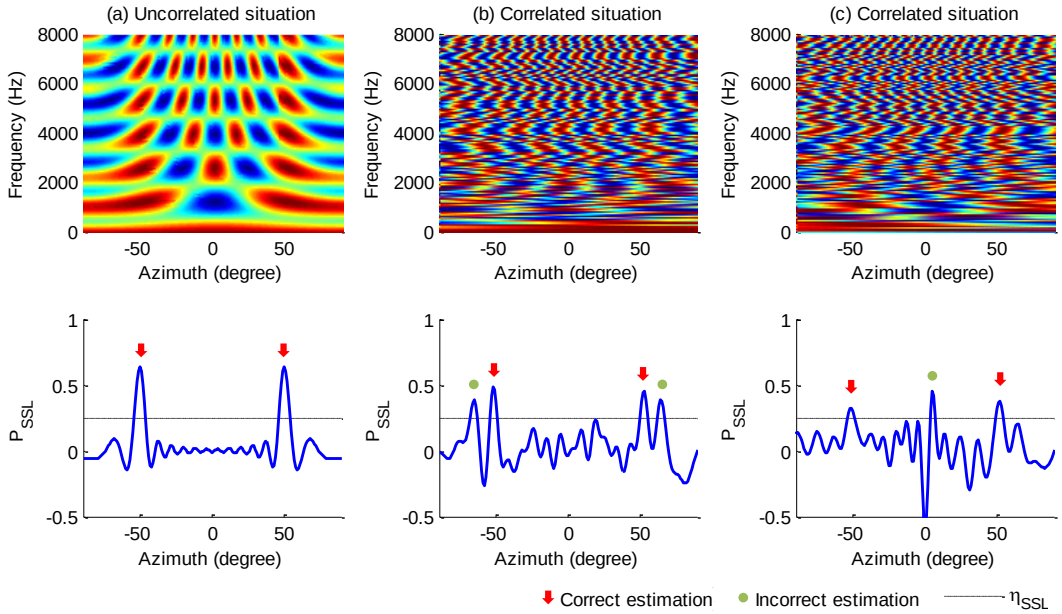

Figure 6.1: Frequency spectrums and peak distributions of ML-based SSL method for two sound sources coming from angles of −50° and +50°: (a) two uncorrelated sources; (b) (c) two highly correlated sources measured in different time frames.

with a threshold η_{SSL} ranging from 0 to 1, as follows:

$$\begin{aligned} &\textit{if} \quad \hat{P}_{SSL}[\theta, n] > \eta_{SSL} \quad \textit{and} \quad \theta \quad \textit{has a peak} \\ &\textit{then} \quad \theta \in \hat{\Theta}_{mle_k}[n] \qquad\qquad , \end{aligned} \tag{6.5}$$

where

$$\hat{P}_{SSL}[\theta, n] = \frac{1}{F}\sum_{f=1}^{F} \frac{X_l[f,n]X_r^*[f,n]}{\left|X_l[f,n]X_r^*[f,n]\right|} \exp\left(j2\pi \frac{f}{F} fs\tau_{multi}(\theta) \right). \tag{6.6}$$

6.3 Difficulties with Multisource sound localization in Real Environments

Multisource sound localization using (6.5) and (6.6) can produce accurate estimates of the sound incidence directions when the sound sources are uncorrelated. However, the

accuracy deteriorates when multiple sound sources are correlated, which is generally the case in real environments, i.e., when the sound sources are speech. For example, if two correlated sound sources are assumed to come from different directions, (6.1) can be rewritten as

$$\begin{aligned} X_l[f,n] = & \\ & \alpha_{l1}[f]\left|S_1[f,n]\right|\exp\left(-j2\pi\frac{f}{F}fs\tau_{l1}\right)+\alpha_{l2}[f]\left|S_2[f,n]\right|\exp\left(-j2\pi\frac{f}{F}fs\tau_{l2}\right)+N_l[f,n] \\ X_r[f,n] = & \\ & \alpha_{r1}[f]\left|S_1[f,n]\right|\exp\left(-j2\pi\frac{f}{F}fs\tau_{r1}\right)+\alpha_{r2}[f]\left|S_2[f,n]\right|\exp\left(-j2\pi\frac{f}{F}fs\tau_{r2}\right)+N_r[f,n], \end{aligned} \tag{6.7}$$

and their cross-power spectrum can be expressed as

$$\begin{aligned} X_l[f,n]X_r^*[f,n] = & \alpha_{l1}[f]\alpha_{r1}[f]\left|S_1[f,n]\right|^2\exp\left(j2\pi\frac{f}{F}fs(\tau_{r1}-\tau_{l1})\right) \\ & +\alpha_{l2}[f]\alpha_{r2}[f]\left|S_2[f,n]\right|^2\exp\left(j2\pi\frac{f}{F}fs(\tau_{r2}-\tau_{l2})\right) \\ & +\alpha_{l1}[f]\alpha_{r2}[f]\left|S_1[f,n]\right|\left|S_2[f,n]\right|\exp\left(j2\pi\frac{f}{F}fs(\tau_{r2}-\tau_{l1})\right) \\ & +\alpha_{l1}[f]\alpha_{r2}[f]\left|S_1[f,n]\right|\left|S_2[f,n]\right|\exp\left(j2\pi\frac{f}{F}fs(\tau_{r1}-\tau_{l2})\right). \end{aligned} \tag{6.8}$$

We can verify that there are two more incorrect TDOAs produced by the correlation between the two sound sources represented in (6.8). Moreover, if we assume a situation in which there are more than two sources or in which additive noise is correlated with other sound sources, the number of incorrect TDOAs will increase geometrically. This phenomenon causes ambiguity in the ML-based SSL method because there will be many peaks in incorrect directions as well in correct ones. Figure 6.1 shows example peak distributions obtained in the ML-based SSL method for two sound signals coming from angles of −50° and +50°. In the uncorrelated situation (a), two sound sources were virtually generated; in the correlated situation (b and c), two speech signals (one for a male and one for a female) recorded at the same time by the SIG-2 humanoid robot were used. When the sound sources were correlated, the ML-based SSL method inaccurately estimated multiple sound incidence directions because of the numerous peaks spread in all directions. In addition, since the intensity of each peak changed over time because

each attenuation factor in (6.8) was applied to each sound source, the ML-based SSL method may select peaks in incorrect directions as correct directions when the peak intensities in the correct directions are lower than those in the incorrect directions. Furthermore, the ML-based SSL method with a threshold frequently fails to produce the same number of directions as sound sources, especially in the absence of information on how many sound sources are active.

These results show that additional functions are needed to minimize the incorrect direction estimations for accurate multisource sound localization.

6.4 Improved Multisource Sound Localization

The solution to the difficulties with multisource sound localization described above is presented here. Approaches to the solution are based on two methods:

- **SNR-weighting function**: if the target sound sources are localized speech in noisy environments, the incorrect direction estimations due to the correlation with additive noise can be prevented by an SNR-weighting function in multisource sound localization.
- **Improved *K*-means clustering**: *K*-means clustering is a commonly used data mining algorithm featuring computational simplicity and high speed. The standard *K*-means clustering algorithm was improved to work well for multisource sound localization in real situations and used it as a tracking method to eliminate incorrect direction estimations due to the correlation between sound sources in each time frame.

6.4.1 SNR-Weighting Function

A spectral weighting function based on the *instantaneous* SNR with the Wiener amplitude estimator was incorporated into the ML-based SSL method to minimize the correlation with additive noise in multisource sound localization.

The *instantaneous* SNR that can be interpreted as a direct estimation of the *a priori* SNR in a spectral subtraction approach [92] is defined by

$$SNR_{inst}[f,n] = \frac{\left|X_l[f,n]X_r^*[f,n]\right| - \mathrm{E}\left[\left|N_l[f,n]N_r^*[f,n]\right|\right]}{\mathrm{E}\left[\left|N_l[f,n]N_r^*[f,n]\right|\right]}. \tag{6.9}$$

Then, the SNR-weighting function for multisource sound localization is obtained by applying (6.9) to the Wiener amplitude estimator as follows:

$$\omega_{SNR}[f,n] = \frac{SNR_{inst}[f,n]}{1+SNR_{inst}[f,n]}. \tag{6.10}$$

The derived SNR-weighting function is incorporated into the multisource sound localization in (6.6) as follows:

$$\hat{P}_{SSL}[\theta,n] = \frac{1}{F}\sum_{f=1}^{F}\omega_{SNR}[f,n]\frac{X_l[f,n]X_r^*[f,n]}{\left|X_l[f,n]X_r^*[f,n]\right|}\exp\left(j2\pi\frac{f}{F}fs\tau_{multi}(\theta)\right). \tag{6.11}$$

6.4.2 Improved *K*-means Clustering

If the directions estimated using (6.5) and (6.11) in the given time frames are the observations to be clustered and if their cluster centers represent the tracked directions for a specific time frame, i.e., given the initial sets of observations (θ_{mle_1}, θ_{mle_2},…, θ_{mle_p}) and *K*-clusters (Θ_{track_1}, Θ_{track_2}, …, Θ_{track_k}) with their center means (θ_{track_1}, θ_{track_2}, …, θ_{track_k}), the standard *K*-means algorithm proceeds by alternating between two steps [93]:

Assignment Step. Assign each observation to the cluster with the closest mean:

$$\Theta_{track_k}^{(i)} = \{\hat{\theta}_{mle_p} : \left|\hat{\theta}_{mle_p} - \theta_{track_k}^{(i)}\right|^2 \le \left|\hat{\theta}_{mle_p} - \theta_{track_j}^{(i)}\right|^2 \ \forall 1 \le j \le K\}, \tag{6.12}$$

where *p* denotes the index of all estimated DOA in the given time frames and *i* denotes the iteration number. Each initial center mean is randomly assigned [94] and each DOA estimation θ_{mle_p} goes into exactly one cluster Θ_{track_k}.

Update Step. Calculate the new means to be the centroid of the observations in each cluster:

$$\theta_{track_k}^{(i+1)} = \frac{1}{\left\langle \Theta_{track_k}^{(i)} \right\rangle}\sum_{\hat{\theta}_{mle_p} \in \Theta_{track_k}^{(i)}} \hat{\theta}_{mle_p}, \tag{6.13}$$

where $<\Theta_{track\_}>_k$ is the number of estimated directions belonging to cluster Θ_{track_k}. These two steps are repeated until the assignments no longer change.

There are two problems with the standard *K*-means clustering when it is to be used for multisource sound localization:

- **Fixed number of clusters**: the number of clusters is fixed from the beginning to the end of the standard *K*-means clustering calculations. This means that the number of sound sources needs to be known in advance for exact clustering. Furthermore, the number of clusters cannot be automatically changed in the observation period for clustering even though sound signals independently appear and disappear over time.
- **Absence of a function for filtering out incorrect direction estimations**: in the standard *K*-means clustering, the tracked directions of the sound signals are not correct because even incorrect direction estimations are used for calculating the center of each cluster.

These two problems cause errors in the results of multisource sound localization. For accurate multisource sound localization, the standard *K*-means clustering was improved by including two additional steps with new criteria:

Increase Step. Increase the number of clusters automatically:

$$\begin{aligned} &\textit{if } \frac{1}{\left\langle \Theta^{(i)}_{track_k} \right\rangle} \sum_{\hat{\theta}_{mle_p} \in \Theta^{(i)}_{track_k}} \left| \hat{\theta}_{mle_p} - \theta^{(i)}_{track_k} \right|^2 > \eta_{C1} \\ &\textit{then } K^{(i+1)} = K^{(i)} + 1 \textit{ and } \textit{move to} \text{ Assignment Step} \\ &\textit{else} \qquad\qquad \textit{move to} \text{ Elimination Step.} \end{aligned} \tag{6.14}$$

The *K*-means clustering algorithm begins with one cluster *(K=1)*. After executing the assignment step and the update step, it adds another cluster *(K=K+1)* if the variance of observations in each cluster is more than a given threshold η_{C1}.

Elimination Step. Eliminate clusters containing incorrect direction estimations:

$$\begin{aligned} &\textit{if } \frac{\left\langle \Theta^{(i)}_{track_k} \right\rangle}{\sum_{k=1}^{K} \left\langle \Theta^{(i)}_{track_k} \right\rangle} < \eta_{C2} \quad \textit{then} \quad \text{eliminate } \textit{cluster } \Theta^{(i)}_{track_k} \\ &\textit{else} \qquad\qquad \text{keep } \textit{cluster } \Theta^{(i)}_{track_k} \quad . \end{aligned} \tag{6.15}$$

The increase step maximizes the number of clusters by using the variance of direction

estimations in each cluster. In this case, some clusters will likely contain few direction estimations that are all incorrect. The elimination step filters out the clusters containing incorrect direction estimations by checking the ratio between the number of direction estimations in each cluster and the number of all direction estimations in the given time frames with a given threshold η_{C2}.

The process of the improved *K*-means clustering algorithm for multisource sound localization is thus as follows:

Step 1. The standard *K*-means algorithm (the assignment step and the update step) is executed with *K=1*.

Step 2. The standard *K*-means algorithm is repeated with *K=K+1* on the basis of Criterion (6.14).

Step 3. All clusters containing incorrect DOA estimations are eliminated on the basis of Criterion (6.15).

6.5 Evaluation

The proposed methods for multisource sound localization were evaluated in two- and three-speaker situations. The subject of the experiment was the SIG-2 humanoid robot. The methods described above were implemented using the HARK open-sourced robot audition software in real time. Figure 6.2 shows the flow of the implemented robot audition system and the process of multisource sound localization. The estimated sound incidence directions were used to make the robot turn at its neck and waist in order to look in the speaker's directions.

6.5.1 Experiments

The experimental were conducted in a room (6 m long × 4.25 m wide × 2.85 m high) with a reverberation time of about 120 ms and noise from air conditioners and personal computers. To create noisier environments, background music with lyrics was played on a laptop just below the robot as additive noise. The average sound pressure levels (SPL) of background music were about 63.29 dB, 67.71 dB, and 72.32 dB, respectively, and those of target speech signals were about 87.3 dB. The SIG-2 humanoid robot was placed at the center of the room and the speakers were located 1.5–2.5 m from the robot.

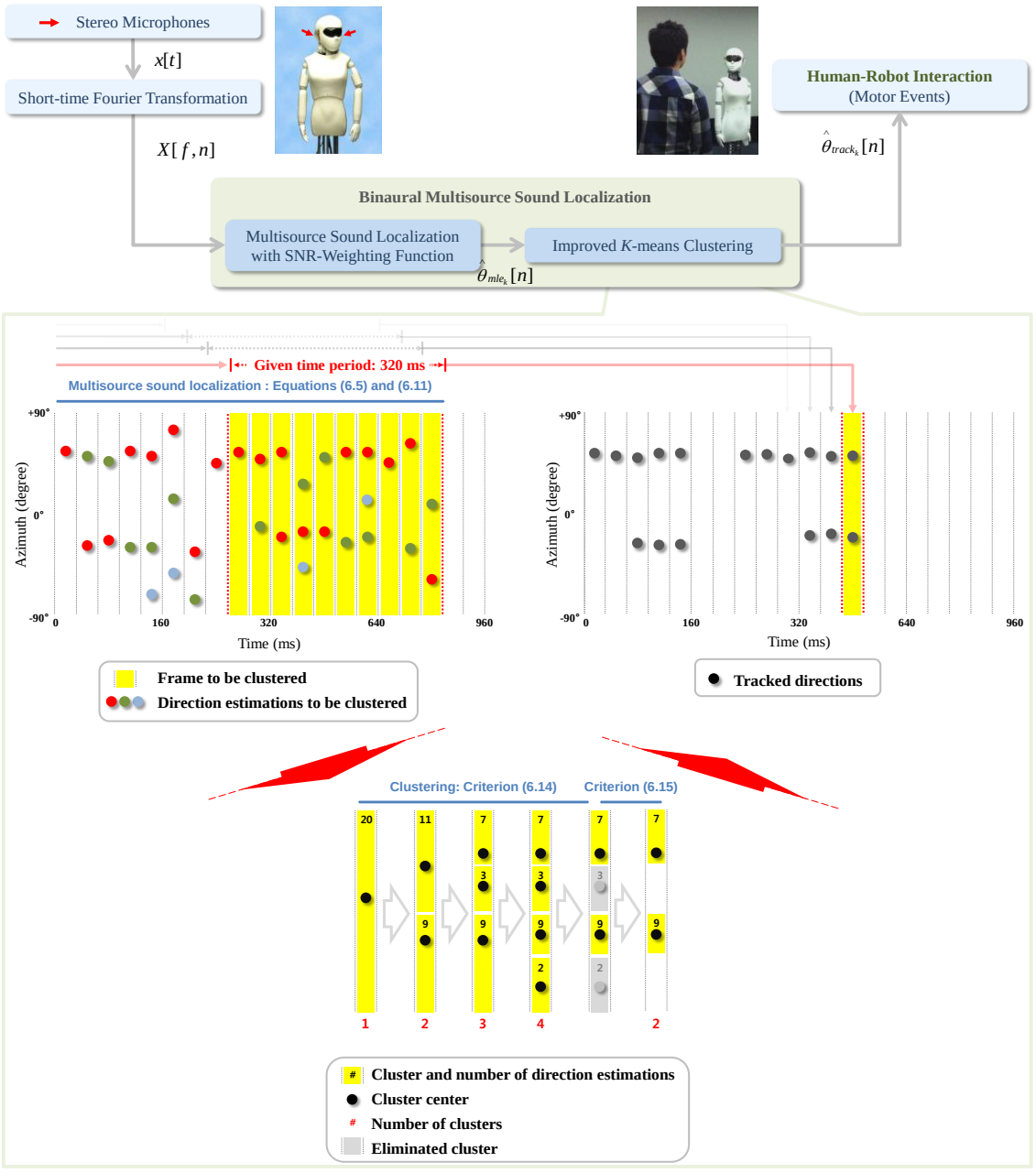

Figure 6.2: Flow of real-time multisource sound localization in SIG-2 humanoid robot.

The thresholds η_{SSL} in (6.5), η_{C1} in (6.14), and η_{C2} in (6.15) were set to 0.23, 3.0, and 0.25, respectively. The system recorded the background noise for 2 s before operating the SIG-2 humanoid robot to estimate the *a priori* noise for the SNR-weighting function. Then two- and three-speaker situations were presented with two males and one female in twelve different conditions of speakers' positions and their speech signals were captured

continuously. The multisource speech localization method was evaluated for each 6 s period.

6.5.2 Experimental Results

The proposed multisource sound localization method consisting of SNR-weighting function and improved *K*-means clustering showed good overall performance even though it sometimes failed in estimating with the exact number of directions.

Tables 6.1 and 6.2 show RMSEs for two- and three-speaker localizations changing the SPL of background music in the twelve different conditions of speakers' positions individually, where each result of RMSEs corresponds to each position of speakers. The RMSE for each estimated direction was less than 5.96° for both the two- and three-speaker situations.

Figure 6.3, 6.4, and 6.5 show the graphical results of two- and three-speaker localization for each 6 s period, where the directions were estimated in 128-ms time frame, 32-ms time shift, and 320-ms time duration for clustering (10 time frames); (a) (b) signal inputs to left and right microphones consisting of male speech signal and female speech signal; (c) actual directions and speech durations of speakers; (d) results of multisource sound localization using (6.5)–(6.6), where colors (red, green, and blue) indicate peaks heights in ascending order and the dotted lines show the actual directions; (e) results of multisource sound localization with the SNR-weighting function using (6.5) and (6.11), where the dotted lines show the actual directions; (f) results of multisource speech localization after performing the improved *K*-means clustering.

Even though the multisource sound localization using (6.5)–(6.6) produced many incorrect direction estimations (shown by (d)), the our multisource sound localization method filtered them out and estimated the direction of each speaker in the running-time domain regardless of changes in the number of speakers over time. The multisource sound localization with SNR-weighting function using (6.5) and (6.11) prevented incorrect direction estimations due to the correlation with additive noise (shown by (e)). The improved *K*-means clustering method filtered out incorrect directions and estimated three speakers regardless of the periods in which they spoke much more accurately than the multisource sound localization using (6.5)–(6.6) (shown by (f)).

Table 6.1: RMSEs of multisource sound localization in two-speaker situations.

Positions of two speakers	SNRs	RMSEs
-90°, +90°	15 dB	2.01°, 3.28°
	20 dB	1.60°, 2.50°
	25 dB	1.43°, 2.96°
-60°, +60°	15 dB	4.58°, 3.87°
	20 dB	3.29°, 2.90°
	25 dB	2.19°, 2.45°
-50°, +30°	15 dB	2.88°, 3.41°
	20 dB	2.68°, 3.82°
	25 dB	2.38°, 3.26°
-20°, +40°	15 dB	2.75°, 4.44°
	20 dB	2.56°, 3.63°
	25 dB	2.09°, 2.57°
-10°, +20°	15 dB	4.69°, 4.62°
	20 dB	3.12°, 4.04°
	25 dB	2.92°, 3.82°
-10°, +30°	15 dB	4.34°, **5.66°**
	20 dB	3.47°, 4.27°
	25 dB	2.27°, 2.35°

The maximum RMSE in two-speaker situations is in bold type.

As a result, despite the use of only two microphones, the robot audition system showed good overall performance for binaural multisource sound localization in real environments.

6.6 Summary

In this chapter, the ML-based SSL method based on the GCC-PHAT method was extended to enable simultaneous multidirection estimations and the difficulties with multisource sound localization using two microphones in a practical situation was described. The difficulties were addressed by incorporating an SNR-weighting function

Table 6.2: RMSEs of multisource sound localization in three-speaker situations.

Positions of three speakers	SNRs	RMSEs
-60°, +30°, +70°	15 dB	4.54°, 3.92°, 2.94°
	20 dB	2.96°, 3.19°, 2.53°
	25 dB	2.23°, 1.82°, 1.34°
-50°, -10°, +30°	15 dB	4.45°, 5.14°, 5.10°
	20 dB	2.71°, **5.96°**, 4.37°
	25 dB	2.25°, 4.50°, 2.43°
-50°, +30°, +60°	15 dB	4.59°, 3.18°, 4.21°
	20 dB	3.10°, 2.01°, 2.73°
	25 dB	1.82°, 1.36°, 1.43°
-30°, +10°, +30°	15 dB	2.71°, 2.39°, 4.36°
	20 dB	1.91°, 2.89°, 3.31°
	25 dB	0.84°, 3.35°, 3.21°
-30°, 0°, +60°	15 dB	3.43°, 3.80°, 5.01°
	20 dB	2.17°, 2.99°, 3.32°
	25 dB	1.47°, 2.62°, 3.63°
-10°, +10°, +30°	15 dB	3.81°, 2.70°, 3.84°
	20 dB	2.44°, 4.68°, 2.77°
	25 dB	2.32°, 3.21°, 1.96°

The maximum RMSE in three-speaker situation is in bold type.

into the ML-based SSL method and performing the *K*-means clustering. To make multisource sound localization more effective in the unknown time-varying number of sound sources situation, the standard *K*-means clustering algorithm was improved by applying two new steps that increase the number of clusters automatically and eliminate clusters containing incorrect direction estimations.

Experimental results demonstrated that the binaural robot audition system with proposed methods can localize directions of multiple sound sources in real time regardless of changes in the number of speakers and periods during which they spoke with localization error below 5.96°. The video demonstration for this work is available on Youtube [95].

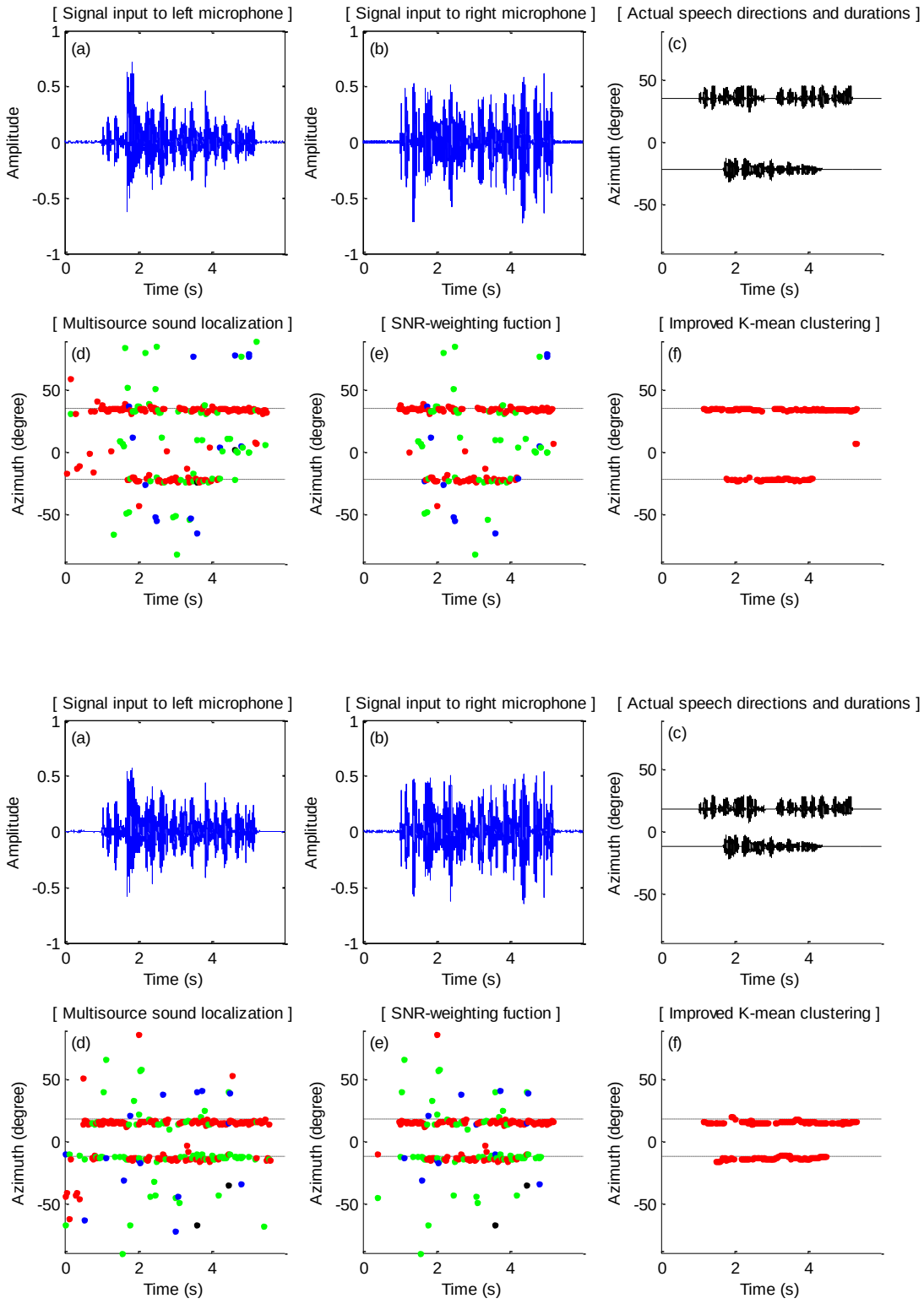

Figure 6.3: Results of two-speaker localization I.

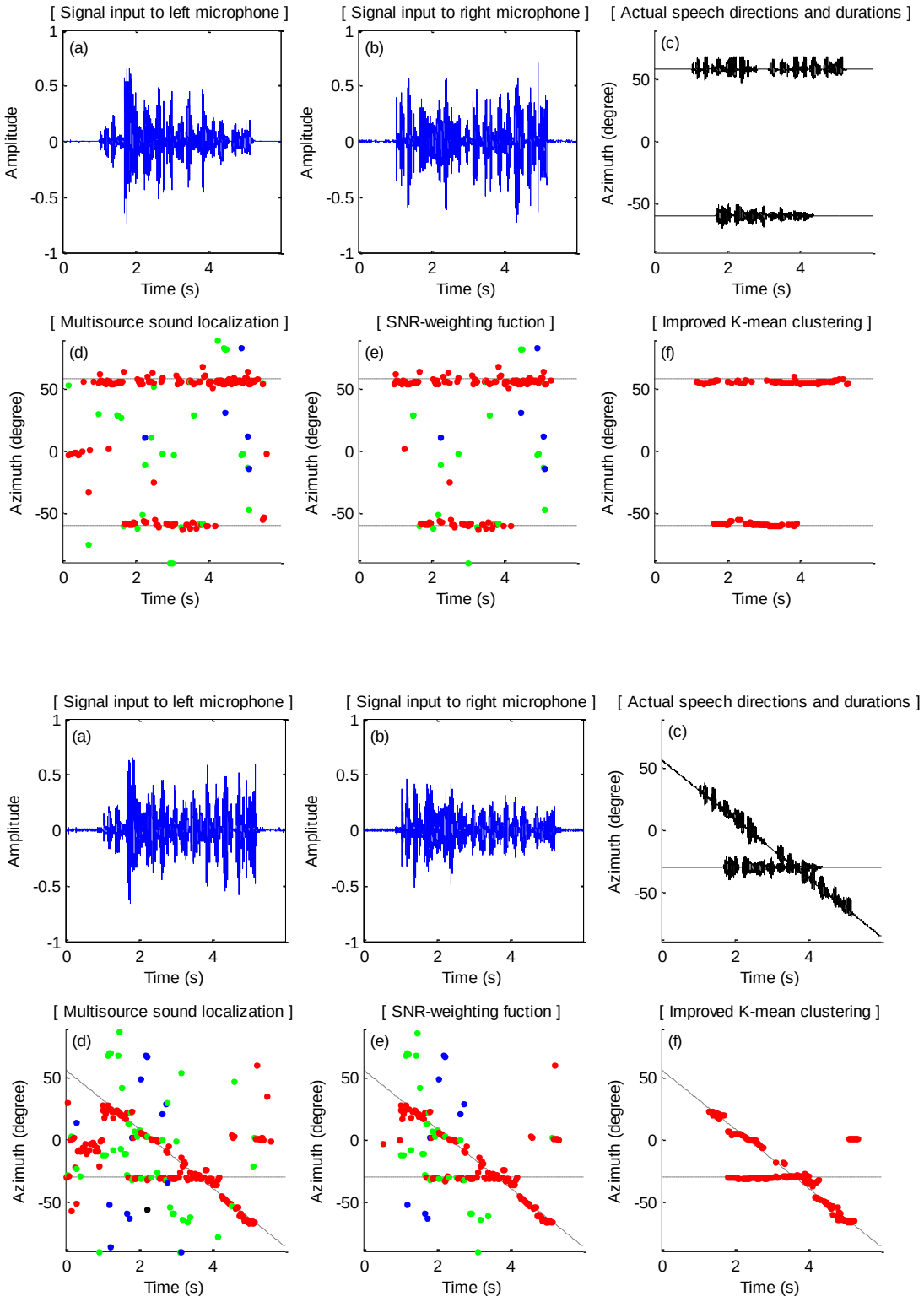

Figure 6.4: Results of two-speaker localization II.

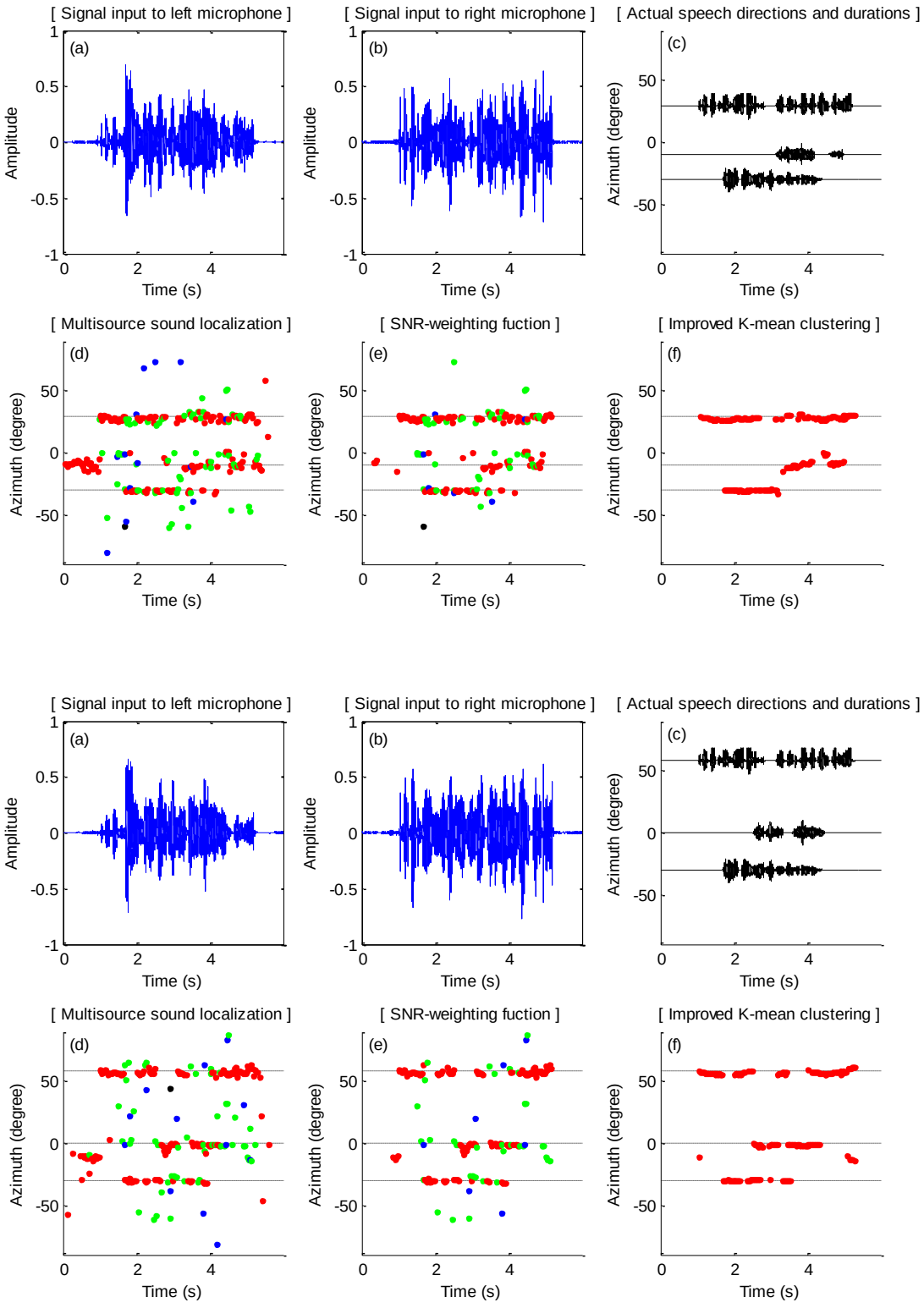

Figure 6.5: Results of three-speaker localization.

CHAPTER 7
Discussion

7.1 Observations

This section summarizes and discusses the main observations from this study, mainly focused on improving technical methods for SSL in binaural robot audition.

7.1.1 Voice Activity Detection for Sound Source Localization

An improved statistical model-based VAD algorithm employing the TSNR technique with recursive noise adaptation was presented for the low SNR cases. In various noisy environments, the computational cost and unexpected SSL errors could be reduced with this improved VAD algorithm by deactivating the SSL process during speech-absent period. The improved VAD algorithm using the TSNR technique with recursive noise adaptation in Chapter 3 showed 5.31 dB better *a priori* SNR estimations during the speech-present period compared to those of the existing statistical model-based VAD algorithm using the power subtraction method. Consequentially, this improved *a priori* SNR estimation made the statistical model-based VAD algorithm to provide a more accurate VAD decision and facilitate more effective SSL in noisy environments as described in Chapters 4 and 5.

7.1.2 Sound Source Localization in Binaural Robot Audition

A sound source localization system based on the GCC-PHAT method for binaural robot audition was developed. Sound signals obtained from only two microphones was used to localize the sound source without any prior information such as HRTF database and parameter settings by training. Two accuracy problems with conventional SSL based on

the GCC-PHAT method in binaural robot audition were addressed: 1) low-resolution TDOA estimation in the time domain and 2) diffraction of sound waves with multipath interference around the robot head. From the experimental results and evaluation, it was verified that the ML-based SSL method in the frequency domain, which was applied as a solution to the first problem, contributed to 1° localization resolution and improved localization accuracy in all-round directions. In addition, a new time delay factor to take into account the second problem improved localization accuracy particularly in the lateral direction (around ±90°). As a result, despite using only two microphones located in the robot head, the real-time SSL system proposed in Chapter 4 showed the overall performance which is as good as other systems utilizing the microphone array consisted of many microphones with SSL errors below 2.23° for static sound sources and 2.54° for 0.20 m/s moving sound sources.

7.1.3 Binaural Sound Localization over Entire Azimuth

The problem of front-back ambiguity in binaural robot audition, which limits the localization range to the front horizontal space, was addressed. To solve this problem, a new front-back disambiguation method was devised by utilizing the pinna amplification effect as a different approach from the existing disambiguation solutions such as use of head movements or HRTF database. This front-back disambiguation method based on the pinna amplification effect was better able to disambiguate the directions of the input sound signals coming from either the front or back compared to the conventional HRTF-based front-back disambiguation method. The success rate in front-back disambiguation using the pinna amplification effect was 92.28% on average over the entire azimuth. As a result, the implementation of the ML-based SSL method proposed in Chapter 4 with the front-back disambiguation method using the pinna amplification effect enabled the binaural robot equipped with silicon human-like pinnae to localize a sound source over the entire azimuth with the average RMSE of 1.96°.

7.1.4 Binaural Localization of Multiple Sound Sources

Since the localization performance generally drops as the number of microphones is reduced, the number of sound sources that the binaural audition system using two

microphones can localize has been limited to a single source in real environments. The difficulty with multisource sound localization is still one significant problem to be overcome in binaural robot audition. The difficulty with multisource sound localization is generally caused by the correlated sound sources, noise, and reverberation, etc. in real environments. This difficulty was analyzed in detail as the correlation problem between highly correlated multiple sources, noise sources, and reverberant signals in a practical situation in Chapter 6. Incorporating a SNR-weighting function into the ML-based SSL method effectively could minimize localization errors due to the correlation with noise in various noisy environments and performing the improved *K*-means clustering could eliminate most SSL errors due to the correlation with sound sources. Experimental results demonstrated that the proposed multisource sound localization system could localize directions of multiple sound sources in real time regardless of changes in the number of speakers and periods during which they spoke with SSL error below 5.96° in various SNR and multiple-speaker situations.

7.2 Contributions

Studies on binaural robot audition can contribute to the development of a user-friendly interface for the robots to appear humanoid or to be perceived to be like human beings. Moreover, it can also contribute to understanding and discovering the human hearing mechanism.

In this thesis, an effect of multipath interference caused by the shape of the robot head in binaural robot audition was identified and it was concluded that binaural localization performance can be significantly improved by considering the diffraction of the sound waves with this multipath interference. This new effect of multipath interference contributes to future binaural robot audition systems as a solution to one of the problems affecting localization accuracy. In the view point of contribution to understanding or discovering the human hearing mechanism, by a new front-back disambiguation method utilizing the pinna amplification effect, it could be discovered that the pinna amplification effect is also a clue to the human auditory system to distinguish sources coming from in the front and back. In addition, a real-time multisource sound localization system that is robust to real environments was realized in binaural robot audition even though using only two microphones is currently

insufficient for the sound processing in robot audition. At least in these achievements, the study on SSL in binaural robot audition in the thesis carries an important contribution.

7.3 Remaining Works

This section discusses remaining works. The techniques that have been proposed in this thesis do not provide complete solutions to the problems of binaural SSL in robot audition. However, they are powerful and can serve as motivation and basis for further works towards more complete systems. In particular, the proposed techniques are versatile since they are based two input sound signals in the time-frequency representation without any prior information. For example, they can be combined and applied with several of the existing techniques or the audition systems: the new time delay factor proposed in Chapter 4 can be used in MUSIC- or beam forming-based methods as an effective steering vector for the microphones array of spherical robot heads; the front-back disambiguation method proposed in Chapter 5 can be utilized for any binaural robot head equipped with two human-like pinnae to distinguish whether the detected sound source is in the front or back; the improved *K*-means clustering proposed in Chapter 6 can be effectively used with any SSL method as a post processing method to filter SSL errors out and smooth direction estimations in real time.

Additionally, four remaining works are mentioned here for extensions that may possibly improve the proposed techniques and lead further steps towards a robust binaural robot audition system.

7.3.1 Multisource Sound Localization with Front-Back Disambiguation

The front-back disambiguation method based on the pinna amplification effect was proposed. However this disambiguation method is not a complete solution to the problem of front-back ambiguity. It still has an ambiguity problem in multisource sound situations. If there are multiple sound sources which has different magnitudes in the front and back respectively, the proposed front-back disambiguation method will fail in distinguish due to ambiguous level differences by mixed intensities of multiple sources

in the same frequency range. For more powerful front-back disambiguation even in multisource sound situations, a blind sound source separation (BSS) technique [96]–[98] may be needed to classify level differences of multiple sources individually to eliminate the ambiguity of level differences between multiple sound sources in the same frequency range.

7.3.2 Multisource Sound Localization with Source Identification

A line of study will be extending the multisource sound localization method with sound source identification [99] so that it can dynamically deal with several moving sound sources. This study focused on a multisource sound localization without sound source identification. Several potential problems happen in situations of multiple moving sound sources, such as incorrect tracking due to ambiguity of speaker identification when moving sound sources cross paths or when they are in the same direction. Humans can localize and tract several moving sound sources even when they are crossing each other and can estimate the exact number of sound sources even in the same direction. This remaining work may be handled by incorporating a sound source identification method to the proposed multisource sound localization method.

7.3.3 Estimating Elevation and Distance of Sound Sources

The human auditory system has capabilities to determine the elevation and distance of sound sources. It is well known that the human auditory system exploits spectral cues (among other ones) mostly generated by the pinnae in order to determine the elevation of a source. For instance, ear asymmetry allows spectral modification for sound originating from below the eye level to sound louder in the left ear, while sound originating from above the eye level to sound louder in the right ear. However, the relation between elevation and spectral cues seems to be quite complex and subjective. This is difficult to exploit in a computational model [100]–[102]. The spectral cues, including the loss of amplitude, the loss of high frequencies, and the ratio of the direct signal to the reverberated signal, are also useful for determining the distance of a source [103]. In this thesis, a part of the spectral modification by the pinnae was investigated

and utilized for front-back disambiguation in the horizontal space. By discovering the relation between elevation or distance and spectral cues, the binaural audition system will be able to localize sound sources in terms of three-dimensional position: the azimuth, the elevation, and the distance for static sound sounds or velocity for moving sound sounds.

7.3.4 Active Robot Audition

Some researchers adopt an approach called as “Active Audition”, which means that the robot moves its microphones embedded on its body [57], [104]–[106]. When humans accurately find the direction of sound sources, the most frequently used head movements involves rotating [107], tipping, and pivoting. Applying the capability of hearing with head movements like human beings into the robot audition system will make binaural SSL performance superior [108]. In this thesis, SSL performance have been improved in stationary situations, i.e., without head movements. As a future work, applying head movements can contribute to the improvements of SSL performance for the robot to be perceived like human beings and understanding human hearing mechanism more deeply.

CHAPTER 8
Conclusion

The goal of this thesis was to improve methods for sound source localization (SSL) in the binaural robot audition using only two microphones inside artificial pinnae like human eras. This was motivated by the performance of the human auditory system and the five technical problems in binaural sound localization, as discussed in the introduction chapter. An overview of existing robot audition systems and several computational techniques in signal processing were given in Chapter 2. With this background knowledge, new localization techniques were proposed as solutions to the five technical problems and the localization performance in binaural robot audition could be significantly improved.

In Chapter 3, a statistical model-based voice activity detection (VAD) algorithm employing the two-step noise reduction (TSNR) technique with recursive noise adaptation was proposed to facilitate SSL processing and reduce unexpected SSL errors during sound-absent period. By improving the *a priori* SNR estimation with the TSNR technique and recursive noise adaptation, it could produce a better VAD results as a significant building block of the SSL system presented in Chapters 4 and 5.

In Chapters 4 and 5, an improved binaural robot audition system was developed that can localize a sound source accurately over the entire azimuth with the average localization errors below 2.23° for static sound sources and 2.54° for 0.20 m/s moving sound sources. For this purpose, the three technical problems with SSL based on the generalized cross-correlation (GCC) method with the phase transform (PHAT) weighting in binaural robot audition were addressed: 1) low-resolution time-delay-of-arrival (TDOA) estimation in the time domain, which makes SSL inaccurate in all directions and impossible in some cases, and 2) diffraction of sound waves with multipath interference around the robot head, which degrade SSL accuracy mostly in the

lateral directions, and 3) front-back ambiguity, which limits the localization range to the front horizontal space. To solve the first problem, the maximum likelihood (ML) estimation was applied to the GCC-PHAT method in the frequency domain instead of the conventional way which uses the cross-power spectrum phase (CSP) analysis. To solve the second problem, a new time delay factor that takes into account the diffraction of sound waves with multipath interference was derived under the assumption that the robot head is spherical and incorporated into the GCC-PHAT method. To solve the third problem, a front-back disambiguation method utilizing the pinna amplification effect was devised. Experimental results demonstrated that applying the frequency-domain ML estimation to the conventional GCC-PHAT method gives 1° SSL resolution and accuracy improvement in all-round directions, that taking the diffraction of sound waves with multipath interference into account is a key to improving localization accuracy in binaural robot audition, and that utilizing the pinna amplification effect successfully solves the problem of front-back ambiguity for robots equipped with two artificial human-like pinnae.

In Chapter 6, the difficulties with multisource sound localization in real environments was addressed by incorporating an SNR-weighting function into the SSL method to minimize noise influence in various noisy environments and performing the *K*-means clustering to eliminate localization errors due to the correlation problem between multiple sources each other in source-correlated environments. For effective multisource sound localization with the unknown time-varying number of sound sources, the standard *K*-means clustering algorithm was improved by applying two new steps that increase the number of sound sources automatically and eliminate clusters including incorrect direction estimations in real time. With the proposed multisource sound localization method, the binaural robot audition system could correctly localize directions of three speakers in real time regardless of changes in the number of speakers and periods during which they spoke with localization error below 5.96°.

List of Publications

International Journals

1. **Ui-Hyun Kim** and Hiroshi G. Okuno, "Improved Binaural Sound Localization and Tracking for Unknown Time-Varying Number of Speakers," *Advanced Robotics*, vol. 27, no. 15, pp. 1161–1173, July 2013.
   → Chapters 4 & 6

2. **Ui-Hyun Kim**, Kazuhiro Nakadai, and Hiroshi G. Okuno, "Improved Sound Source Localization and Front-Back Disambiguation for Humanoid Robots with Two Ears," *Lecture Notes in Computer Science, Recent Trends in Applied Artificial Intelligence, Springer-Verlag Berlin Heidelberg*, vol. 7906, pp. 282–291, received **The Best Paper Award**, June 2013.
   → Chapters 4 & 5

International Conferences

1. **Ui-Hyun Kim** and Hiroshi G. Okuno, "Robust Localization and Tracking of Multiple Speakers in Real Environments for Binaural Robot Audition," in *Proceedings of the International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIA^{2}MIS)*, no. 40, Paris, France, July 2013.
   → Chapter 6

2. Randy Gomez, Keisuke Nakamura, Shinsuke Mori, Kazuhiro Nakadai, **Ui-Hyun Kim**, Hiroshi G. Okuno, and Tatsuya Kawahara, "Hands-Free Human Robot Communication Robust to Speaker's Radial Position," in *Proceedings of the IEEE International Conference on Robots and Automation (ICRA)*, pp. 4314–4319, Karlsruhe, Germany, May 2013.

3. **Ui-Hyun Kim**, Takeshi Mizumoto, Tetsuya Ogata, and Hiroshi G. Okuno, "Improvement of Speaker Localization by Considering Multipath Interference of Sound Wave for Binaural Robot Audition," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2910–2915, San Francisco, USA, September 2011.
   → Chapter 4

National Conferences

1. **Ui-Hyun Kim** and Hiroshi G. Okuno, "Reliable Speaker Localization using Signal-to-Noise Ratio Information in Noise Environments for the SIG-2 Humanoid Robot," in *Proceedings of the Annual Conference of the Robotics Society of Japan (RSJ)*, 4D1-4, Sapporo, Japan, September 2012.
   → **Chapter 6**

2. **Ui-Hyun Kim**, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno, "Improved Statistical Model-Based Voice Activity Detection with Noise Reduction for the SIG-2 Humanoid Robot," in *Proceedings of the Annual Conference of the Robotics Society of Japan (RSJ)*, 1Q1-7, Tokyo, Japan, September 2011.
   → **Chapter 3**

3. **Ui-Hyun Kim**, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno, "Speaker Localization Using Two-Channel Microphone on the SIG-2 Humanoid Robot," in *Proceedings of the National Convention of Information Processing Society of Japan (IPSJ)*, 4C-1, Tokyo, Japan, March 2011.
   → **Chapter 4**

Bibliography

[1] J. Casper and R. Murphy, "Human-robot interaction during the robot-assisted urban search and rescue effort at the world trade center," *IEEE Trans. Systems. Man, and Cybernetics-Part B: Cybernetics*, vol. 33, No. 3, pp. 367–385, June 2003.

[2] K. Dautenhahn, "Socially intelligent robots: Dimensions of human-robot interaction," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 362, no. 1480, pp. 679–704, April 2007.

[3] Daniel Starch. *Perimetry of the localization of sound*. State University of Iowa. 1908.

[4] J. M Valin, F. Michaud, J. Rouat, and D. Letouneau, "Robust Sound Source Localization Using a Microphone Array on a Mobile Robot," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp. 1128–1233, Las Vegas, USA, October 2003.

[5] Y. Tamai, Y, Sasaki, S. Kagami, and H. mizoguchi, "Three Ring Microphone Array for 3D Sound Localization and Separation for Mobile Robot Audition," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp. 4172–4177, Alberta, Canada, August 2005.

[6] U. H. Kim, J. Kim, D. Kim, H. Kim, and B. J. You, "Speaker Localization Using the TDOA-based Feature Matrix for a Humanoid Robot," in *Proc. IEEE Int. Symp. on Robot and Human Interactive Communication (RO-MAN)*, pp. 610–615, Munich, Germany, August 2008.

[7] J. S. Hu, C. Y. Chan, C. K. Wang, and C. C. Wang, "Simultaneous Localization of Mobile Robot and Multiple Sound Sources Using Microphone Array," in *Proc. IEEE Int. Conf. on Robots and Automation (ICRA)*, pp. 29–34, Kobe, Japan, May 2009.

[8] H. Sun, P. Yang, Z. Liu, L. Zu, and Q. Xu, "An Auditory System of Robot for Sound Source Localization Based on Microphone Array," in *Proc. IEEE/RSJ Int. Conf. on Robotics and Biomimetics (ROBIO)*, pp. 629–632, Tianjin, China, December 2010.

[9] X. Li, H. Liu, and X. Yang, "Sound Source Localization for Mobile Robot Based on Time Difference Feature and Space Grid Matching," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp. 2879–2886, San Francisco, USA, September 2011.

[10] Y. Sasaki, M. Kabasawa, S. Thompson, S. Kagami, and K. Oro, "Spherical Microphone Array for Spatial Sound Localization for a Mobile Robot," in *Proc. IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems (IROS)*, pp. 713–718, Algarve, Portugal, October 2012.

[11] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization (Revised Edition)*. Cambridge. MA: MIT Press. 1997.

[12] M. Jeub, C. Herglotz, C. Nelke, C. Beaugeant, and P. Vary, "Noise reduction for dual-microphone mobile phones exploiting power level differences," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1693–1696, Kyoto, Japan, March 2012.

[13] H. Wallach, E. B. Newman, and M. R. Rosenzweig, "The precedence effect in sound localization," *American Journal of Psychology*, vol. 62, no. 3, pp. 315–336, July 1949.

[14] J. Blauert and J. Braasch, "Binaural Signal Processing," in *Proc. IEEE Int. Conf. on Digital Signal Processing (DSP)*, pp. 1–11, Greece, July 2011.

[15] T. Rodemann, "A study on distance estimation in binaural sound localization," in *Proc. IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems (IROS)*, pp. 425–430, Offenbach, Germany, October 2010.

[16] K. Youssef, S. Argentieri, and J. L. Zarader, "Towards a Systematic Study of Binaural Cues," in *Proc. IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems (IROS)*, pp. 1004–1009, Vilamoura, Portugal, October 2012.

[17] C. I. Cheng and G. H. Wakefield, "Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space," *Audio Engineering Society*, vol. 49, pp. 231–249, April 2001.

[18] G. C. Carter, A. A. Nuttall, and P. G. Cable, "The Smoothed Coherence Transform," in *Proc. IEEE*, vol. 61, no. 10, pp. 1497–1498, October 1973.

[19] C. H. Knapp and G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[20] Y. Rui and D. Florencio, "Time Delay Estimation in the Presence of Correlated Noise and Reverberation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 133–136, Montreal, Canada, May 2004.

[21] V. M. Trifa, A. Koene, J. Moren, and G. Cheng, "Real-time Acoustic Source Localization in Noisy Environments for Human-robot Multimodal Interaction," in *Proc. IEEE Int. Symp. on Robot and Human Interactive Communication (RO-MAN)*, pp. 393-398, Jeju, Korea, August 2007.

[22] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, January 1999.

[23] J. H. Chang, N. S. Kim, and S. K. Mitra, "Voice Activity Detection Based on Multiple Statistical Models," *IEEE Trans. on Signal Processing*, vol. 54, no. 6, pp. 1965–1976, June 2006.

[24] H. Wallach, "The role of head movements and vestibular and visual cues in sound localization," *Journal of Experimental Psychology*, vol. 27, no. 4, pp. 339–368, October 1940.

[25] P. A. Hill, P. A. Nelson, and O. Kirkeby, H. Hamada, "Resolution of Front-Back Confusion in Virtual Acoustic Imaging Systems," *Acoustical Society of America*, vol. 108, no. 6, pp. 2901–2910, December 2000.

[26] H. Nakashima and T. Mukai, "3D Sound Source Localization System Based on Learning of Binaural Hearing," in *Proc. IEEE Inter. Conf. on Systems, Man and Cybernetics (SMC)*, vol. 4, pp. 3534–3539, Nagoya, Japan, October 2005.

[27] A. Ovcharenko, S. J. Cho, and U. P. Chonga, "Front-back confusion resolution in three-dimensional sound localization using databases built with a dummy head," *Acoustical Society of America*, vol. 122, no. 1, pp. 489–495, August 2007.

[28] T. Rodemann, G. Ince, F. Joublin, and C. Goerick, "Using Binaural and Spectral Cues for Azimuth and Elevation Localization," in *Proc. IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems (IROS)*, pp. 2185–2190, Nice, France, September 2008.

[29] I. Toshima and S. Aoki, "Sound Localization During Head Movement Using an Acoustical Telepresence Robot: TeleHead," *Advanced Robotics*, vol. 23, no. 3, pp. 289–304, 2009.

[30] M. P. Hofman, J. G. V. Riswick, and A. J. V. Opstal, "Relearning sound localization with new ears," *Nature Neuroscience*, vol. 1, no. 5, pp. 417–421. September 1998.

[31] R. H. Y. So, B. Ngan, A. Horner, K. L. Leung, J. Braasch, and J. Blauert, "Toward orthogonal non-individualized head-related transfer functions for forward and backward directional sound: cluster analysis and an experimental study. Ergonomics," *Ergonomics*, vol. 53, no. 6, pp. 767–781, June 2010.

[32] H. D. Kim, K. Komatani, T. Ogata, H. G. Okuno, "Binaural Active Audition for Humanoid Robots to Localise Speech over Entire Azimuth Range," *Applied Bionics and Biomechanics, Special Issue: Humanoid Robots*, vol. 6, no. 3-4, pp. 355–368, November 2009.

[33] H. D. Kim, K. Komatani, T. Ogata, H. G. Okuno, "Human Tracking System Integrating Sound and Face Localization Using an Expectation-Maximization Algorithm in Real Environments," *Advanced Robotics*, vol. 23, no. 6, pp. 629–653, May 2009.

[34] R. O. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Trans. on Antennas Propagation*, vol. AP-34, pp.276–280, March 1986.

[35] H. Viste and G. Evangelista, "Binaural Source Localization," International Conference on Digital Audio Effects (DAFx), pp. 145–150, Italy, October 2004.

[36] K Youssef, S Argentieri, and J. L. Zarader, "A Binaural Sound Source Localization Method Using Auditive Cues and Vision," in *Proc. IEEE Int. Symp. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 217–220, Kyoto, Japan, March 2012.

[37] K. Nakadai, H. G. Okuno, H. Nakajima, and Y. Hasegawa, "An Open Source Software System for Robot Audition HARK and Its Evaluation," in *Proc. IEEE/RSJ Inter. Conf. on Humanoid Robots (Humanoids)*, pp. 561–566, Daejeon, Korea, December 2008.

[38] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and Implementation of Robot Audition System 'HARK' - Open Source Software for Listening to Three Simultaneous Speakers," *Advanced Robotics*, vol. 24, pp. 739–761, 2010.

[39] T. Mizumoto, K. Nakadai, T. Yoshida, R. Takeda, T. Otsuka, T. Takahashi, and H. G. Okuno, "Design and implementation of selectable sound separation on the Texai telepresence system using HARK," in *Proc. IEEE Int. Conf. on Robots and Automation (ICRA)*, pp. 2130–2137, Shanghai, China, May 2011.

[40] C. Plapous, C. Marro, P. Scalart, and L. Mauuary, "A Two-Step Noise Reduction Technique," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 289–292, Montral, Canada, May 2004.

[41] C. Plapous, C. Marro, and P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement", *IEEE Trans. on Audio, Speech & Language Processing*, pp. 2098–2108, 2006.

[42] M. Omologo and P. Svaizer, "Acoustic Event Localization Using A Crosspower-Spectrum Phase Based Technique," in *Proc. IEEE Int. Symp. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 273–276, Adelaide, Australia, April 1994.

[43] J. A. Hartigan, M. A. Wong, "Algorithm AS 136: A *K*-Means Clustering Algorithm," *Journal of the Royal Statistical Society: Series C*, vol. 28, no. 1, pp. 100–108, 1979.

[44] K. Nakadai, H. G. Okuno, and H. Kitano, "Epipolar Geometry Based Sound Localization and Extraction for Humanoid Audition," in *Proc. IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems* (IROS), vol. 3, pp. 1395–1401, Hawaii, USA, October 2001.

[45] H. G. Okuno, K. Nakadai, K. Hidai, H. Mizoguchi, and H. Kitano, "Human-Robot Interaction through Real-Time Auditory and Visual Multiple-Talker Tracking," in *Proc. IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems* (IROS), vol. 3, pp. 1402–1409, Hawaii, USA, October 2001.

[46] K. Nakadai, K. Hidai, H. G. Okuno, and H. Kitano, "Real-Time Speaker Localization and Speech Separation by Audio-Visual Integration," in *Proc. IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems (IROS)*, pp. 1043–1049, Washington DC, USA, May 2002.

[47] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Kitano, "Applying Scattering Theory to Robot Audition System: Robust Sound Source Localization and Extraction," in *Proc. IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems (IROS)*, pp. 1147–1152, Las Vegas, USA, October 2003.

[48] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Tsujino, "Improvement of Recognition of Simultaneous Speech Signals Using AV Integration and Scattering Theory for Humanoid Robots," *Speech Communication*, vol. 44, pp. 97–112, October 2004.

[49] R. G. Newton. *Scattering theory of waves and particles*. Courier Dover Publications. 1982.

[50] H. D. Kim, J. S. Choi, and M. Kim, "Human-Robot Interaction in Real Environments by Audio-Visual Integration", *International Journal of Control, Automation, and Systems*, vol. 5, no. 1, pp. 61–69, February 2007.

[51] H. D. Kim, K. Komatani, T. Ogata, and H. G. Okuno, "Auditory and Visual Integration based Localization and Tracking of Multiple Moving Sounds in Daily-life Environments," in *Proc. IEEE Int. Symp. on Robot and Human Interactive Communication (RO-MAN)*, pp.399-404, Jeju, Korea, August 2007.

[52] H. D. Kim, K. Komatani, T. Ogata, and H. G. Okuno, "Evaluation of Two-Channel-Based Sound Source Localization using 3D Moving Sound Creation Tool," in *Proc. IEEE Int. Conf. on Informatics Education and Research for Knowledge-Circulating Society (ICKS-2008)*, pp.209–212, Kyoto, Japan, January 2008.

[53] H. D. Kim, K. Komatani, T. Ogata, and H. G. Okuno, "Two-Channel-Based Voice Activity Detection for Humanoid Robots in Noisy Home Environments," in *Proc. IEEE Int. Conf. on Robots and Automation (ICRA)*, pp. 3495–3501, California, USA, May 2008.

[54] H. D. Kim, J. Kim, K. Komatani, T. Ogata, and H. G. Okuno, "Target Speech Detection and Separation for Humanoid Robot in Sparse Dialogue with Noisy Home Environments," in *Proc. IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems (IROS)*, pp. 1705–1711, Nice, France, September 2008.

[55] H. D. Kim, K. Komatani, T. Ogata, and H. G. Okuno, "Design and Evaluation of Two-Channel-Based Sound Source Localization over Entire Azimuth Range for Moving Talkers," in *Proc. IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems (IROS)*, pp. 2197–2203, Nice, France, September 2008.

[56] H. D. Kim, K. Komatani, T. Ogata, and H. G. Okuno, "Real-Time Auditory and Visual Talker Tracking through integrating EM algorithm and Particle Filter," *New Trends in Applied Artificial Intelligence, Lecture Notes in Artificial Intelligence*, vol. 4570, pp. 280–290, Springer-Verlag, June 2007.

[57] H. D. Kim, "Binaural Active Audition for Humanoid Robots," *Ph. D. thesis*, Kyoto University, Japan, September 2008.

[58] Okuno Laboratory. *HARK: HRI-JP Audition for Robots with Kyoto University*. Available: http://winnie.kuis.kyoto-u.ac.jp/HARK/

[59] A. S. Bregman. *Auditory scene analysis*, Cambridge. MA: MIT Press. 1990.

[60] S. Yamamoto, K. Nakadai, M. Nakano, H. Tsujino, J. M. Valin, K. Komatani, T. Ogata, and H. G. Okuno, "Real-Time Robot Audition System That Recognizes Simultaneous Speech in The Real World," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp. 5333–5338, Beijing, China, October 2006.

[61] K. Nakadai, G. Ince, K. Nakamura, and H. Nakajima, “Robot Audition for Dynamic Environments,” in *Proc. IEEE Inter. Conf. on Signal Processing, Communication and Computing (ICSPCC)*, pp. 125–130, Hong Kong, China, August 2012.

[62] H. G. Okuno and K. Nakadai, “Computational Auditory Scene Analysis and its Application to Robot Audition,” in *Proc. IEEE Inter. Conf. on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, pp. 124–127, Trento, Italy, May 2008.

[63] P. Schleich, P. Nopp, and P. D’Haese, “Head shadow, squelch, and summation effects in bilateral users of the MED-EL COMBI 40/40+ cochlear implant,” *Ear and Hearing*, vol. 25, pp. 197–204, 2004.

[64] M. M. Van Wanrooij and A. J. Van Opstal, “Contribution of head shadow and pinna cues to chronic monaural sound localization,” *Journal of Neuroscience*, vol. 24, pp. 4163–4171, 2004.

[65] B. C. J. Moore. *An introduction to the psychology of hearing (fifth edition)*. Academic Press. 2003.

[66] D. Wang and G. J. Brown. *Computational auditory scene analysis: principles, algorithms and applications*. Wiley interscience. 2006.

[67] A. D. Musicant, J. C. Chan, and J. E. Hind, “Direction-dependent spectral properties of cat external ear: new data and cross-species comparisons,” *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 757–781, February 1990.

[68] R. H. Y. So, N. M. Leung, J. Braasch, and K. L. Leung, “A low cost, Non-individualized surround sound system based upon head-related transfer functions. An Ergonomics study and prototype development,” *Applied Ergonomics*, vol. 37, pp. 695–707, 2006.

[69] B. D. V. Veen and K.M. Buckley, “Beamforming: A versatile approach to spatial filtering,” *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, April 1988.

[70] H. L. Van Trees. *Optimum Array Processing*. Wiley. NY. 2002.

[71] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*, John Wiley & Sons Inc., 1996.

[72] J. C. Hassab, and R. E. Boucher, “Optimum Estimation of Time Delay by a Generalized Correlator,” *IEEE T-ASSP*, vol. 27, no. 4, pp. 373–380, August 1979.

[73] J. C. Hassab and R. E. Boucher, "Performance of the Generalized Cross Correlator in the Presence of a Strong Spectral Peak in the Signal," *IEEE T-ASSP*, vol. 29, no. 3, pp. 549–555, June 1981.

[74] M. Azaria and D. Hertz, "Time delay estimation by generalized cross correlation methods," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 280–285, April 1984.

[75] S. M. Griebel and M. S. Brandstein, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proc. IEEE Inter. Work. on Applications of Signal Processing to Audio and Acoustic (WASPAA)*, pp. 71–74, New Platz, New York, Ooctorber 2001.

[76] Jont B. Allen, "Short Time Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, June 1977.

[77] M. H. Hayes. *Digital Signal Processing*. Schaum's Outline Series. New York: McGraw Hill. 1999.

[78] J. Ramírez, J. M. Górriz, and J. C. Segura, "Voice Activity Detection. Fundamentals and Speech Recognition System Robustness," in *Robust Speech Recognition and Understanding*, Book edited by M. Grimm and K. Kroschel, pp. 1–22, 2007.

[79] A. Benyassine, E. Shlomot, H.Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit, "ITU-T Recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, vol. 35, pp. 64–73, September 1997.

[80] F. Beritell, S. Casale, G. Ruggeri, and S. Serrano, "Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 85–88, March 2002.

[81] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proc. IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 365–368, May 1998.

[82] Y. Ephraïm and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech, Signal Processing*, vol. assp-32, no. 6, pp. 1109–1121, Dec. 1984.

[83] O. Cappé, “Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor,” *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, April 1994.

[84] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

[85] I. Cohen and B. Berdugo, “Noise estimation by minima controlled recursive averaging for robust speech enhancement,” *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, January 2002.

[86] M. Matassoni and P. Svaizer, “Efficient time delay estimation based on cross-power spectrum phase,” *European Signal Processing Conference (EUSIPCO)*, Florence, Italy, September 2006.

[87] M. Jian, A. C. Kot, M. H. Er, “DOA Estimation of Speech Source with Microphone Arrays,” in *Proc. IEEE Int. Symp. on Circuits and Systems (ISCAS)*, pp. 293-296, vol. 5l, Monterey, California, June 1998.

[88] NaturalPoint, Inc. *OptiTrack: 3D Motion capture systems and software*. Available: http://www.naturalpoint.com/optitrack/

[89] J. C. Middlebrooks, “Sound Localization by Human Listeners,” *Annual Review of Psychology*, vol. 42, pp. 135–159, February 1991.

[90] Y. Suzuki, F. Asano, H. Y. Kim, and Toshio Sone, “An Optimum Computer-Generated Pulse Signal Suitable for the Measurement of very Long Impulse Responses,” *Acoustical Society of America*, vol. 97, no. 2, pp. 1119–1123, February 1995.

[91] U. H. Kim, “Improved Sound Source Localization and Front-Back Disambiguation for Humanoid Robots with Two Ears,” Video Demonstration, March 2012. Available: http://youtu.be/iCE--ir-JRc

[92] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwith compression of noisy speech,” *in Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, December 1979.

[93] D. MacKay. *Chapter 20. An Example Inference Task: Clustering*. Information Theory, Inference and Learning Algorithms, Cambridge University Press, pp. 284–292, 2003.

[94] G. Hamerly and C. Elkan, "Alternatives to the *K*-means algorithm that find better clusterings," in *Proc. IEEE Inter. Conf. on Information and knowledge management (CIKM)*, pp. 600–607, Virginia, USA, 2002.

[95] U. H. Kim, "Improved Binaural Sound Localization and Tracking for Unknown Time-Varying Number of Speakers," Video Demonstration, May 2012. Available: http://youtu.be/VYJba0dia7E

[96] P. Comon, "Independent Component Analysis: a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[97] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Publications. New York. 2001.

[98] T. Kim, H. Attias, S. Y. Lee, and T. W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 70–79, January 2007.

[99] A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 804–816, November 2003.

[100] S. R. Oldfield and S. P. Parker, "Acuity of sound localization: a topography of auditory space. II. Pinna cues absent," *Perception*, vol. 13, pp. 601–617, 1984.

[101] R. A. Butler and R. A. Humanski, "Localization of sound in the vertical plane with and without high-frequency spectral cues," *Perception & Psychophysics*, vol. 51, no. 2, pp. 182–186, March 1992.

[102] J. C. Middlebrooks, "Narrow-band sound localization related to external ear acoustics," *Journal of the Acoustical Society of America*, vol. 61, pp. 2607–2624, November 1992.

[103] C. Roads. *The Computer Music Tutorial*. Cambridge. MA: MIT. 2007.

[104] K. Nakadai, T. Matsui, H. G. Okuno, and H. Kitano, "Active Audition System And Humanoid Exterior Design," in *Proc. IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems (IROS)*, vol.2, pp. 1453–1461, Takamatsu, Japan, November 2000.

[105] K. Nakadai, H. Okuno, and H. Kitano, "Robot recognizes three simultaneous speech by active audition," in *Proc. IEEE Int. Conf. on Robots and Automation (ICRA)*, vol. 1, pp. 398–405, May 2003.

[106] E. Berglund and J. Sitte, “Sound source localisation through active audition,” in *Proc. IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems (IROS)*, pp. 653–658, August 2005.

[107] C. F. Altmann, E. Wilczek, and J. Kaiser, “Processing of auditory location changes after horizontal head rotation,” *The Journal of Neuroscience*, vol. 29, no. 41, pp. 13074–13078, October 2009.

[108] I. Toshima and S. Aoki, “The effect of head movement on sound localization in an acoustical telepresence robot: TeleHead,” in *Proc. IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems (IROS)*, pp. 872–877, Beijing, China, October 2006.